

Proceedings in Operations Research

Edited by

Herausgegeben von

M. Henke

A. Jaeger

R. Wartmann

H. J. Zimmermann

Papers of the Annual Meeting

Vorträge der Jahrestagung

DGU

1971



Physica - Verlag

Würzburg - Wien

Proceedings in Operations Research

Herausgegeben

von

M. Henke, A. Jaeger, R. Wartmann-und H.-J. Zimmermann

Vorträge der Jahrestagung 1971

DGU

Papers of the Annual Meeting 1971



Physica-Verlag · Würzburg - Wien

1972

ISBN 3 7908 0119 4

Das Buch oder Teile davon dürfen weder photomechanisch, elektronisch noch in irgendeiner anderen Form
ohne schriftliche Genehmigung des Verlags wiedergegeben werden

©

Physica-Verlag, Rudolf Liebing KG, Würzburg 1972

ISBN-13: 978-3-7908-0119-4 e-ISBN-13: 978-3-642-99745-7
DOI: 10.1007/978-3-642-99745-7

VORWORT

Die 10. Jahrestagung der Deutschen Gesellschaft für Unternehmensforschung zeichnete sich in verschiedenen Hinsichten besonders aus: Sie war nicht nur eine Jubiläumstagung, sondern sie war gleichzeitig die letzte Jahrestagung der DGU. Man könnte geneigt sein, diese Tatsache zu bedauern, wenn nicht gleichzeitig mit der Auflösung der DGU etwas geschehen wäre, was viele deutsche Unternehmensforscher seit geraumer Zeit gewünscht hatten: Die DGU schloss sich mit der Gesellschaft für Operations Research (früher AKOR) zu einer einheitlichen deutschen Gesellschaft, der Deutschen Gesellschaft für Operations Research (DGOR) zusammen. So ist das Ende der DGU gleichzeitig der Beginn der DGOR, einer Gesellschaft, der man nur wünschen kann, dass sie in Fortsetzung der Arbeit der Deutschen Gesellschaft für Unternehmensforschung und der Gesellschaft für Operations Research der Fortschritt auf dem Gebiete des Operations Research fördern möge.

Die zehn Jahre des Bestehens der DGU bezeichnen in Deutschland die Entwicklungsphase des Operations Research oder der Unternehmensforschung, wie sie weitgehend in Deutschland genannt wurde, von den absoluten Anfängen bis zu dem Stadium in dem in weiten Kreisen der Wirtschaft und der Lehre die Nützlichkeit der Anwendung dieser Methoden nicht mehr bezweifelt wird. Dies kommt auch im Programm dieser Jubiläumstagung zum Ausdruck:

Zum einen waren die Themenkreise noch nie so weit gespannt wie in diesem Jahr. Sie reichen von praktischen Anwendungen auf den verschiedensten Gebieten einschliesslich der Politik, des Umweltschutzes und der Familienplanung auf der einen Seite bis zu bedeutenden theoretischen Ergebnissen, von denen noch nicht vorauszusehen ist, auf welchen Wissensgebieten sie einmal angewandt werden.

Zum anderen konnte bisher noch nie eine so hohe internationale Beteiligung verzeichnet werden wie bei dieser Tagung. Neben Unternehmensforschern aus Nachbarländern wie Belgien, den Niederlanden, Österreich und der Schweiz, konnten wir auch Kollegen aus Grossbritannien, den USA und Kanada begrüßen.

Um sowohl den Teilnehmern an der Tagung als auch den Interessierten, die nicht an der Tagung teilnehmen konnten, Gelegenheit zu geben, sich mit den Ergebnissen der Tagung vertraut zu machen, wurden die Tagungsbeiträge in den vorliegenden "Proceedings" veröffentlicht. Dieser Brauch, von der DGU auf ihrer letzten Jahrestagung begonnen, soll ebenfalls von der DGOR weitergeführt werden. Möge er Praktikern und Forschern gleichermassen nützen.

Prof. Dr. H. -J. Zimmermann
(1. Vorsitzender der DGZ)

INHALTSVERZEICHNIS

Einführungen

KÜNZI, H.: Unternehmensforschung in Wissenschaft, Wirtschaft und Politik	3
JAEGER, A.: Beziehungen zwischen Mathematik und Wirklichkeit in mathematischen Modellen von empirischen Strukturen	15
KORTANEK, K.O., and W.L. GORR: Numerical Aspects of Pollution Abatement Problems: Optimal Control Strategies for Air Quality Standards	34
RUTSCH, M.: Entscheidungskriterien - Hilfe oder Hindernis beim Handeln?	59

Kommunale Anwendungen

MEISE, J., und M. WEGENER: Digitale Simulation der räumlichen Stadtentwicklung	81
LIESENFELD, K.P.: Ein Modell zur optimalen Wärmeversorgung in Städten durch leitungsgebundene Energieträger	102

Stochastische Entscheidungsprobleme

RABUSSEAU, R., und W. REICH: Entscheidungen einer Person unter Unsicherheit	123
SCHICK, G.J., and T.M. DRNAS: Bayesian Reliability Demonstration	142
RÖDDER, W.: Lösungsvorschläge für stochastische Zielprogramme	166
SCHNEEWEISS, Ch.: Zur Theilschen Theorie dynamischer Sicherheitsäquivalente	177
DÜRR, W.: Stochastische Programmierungsmodelle als Vektormaximumprobleme	189

Langfristige Planung

PETERS, L.: Ein Modell für die strategische Zielsetzung und Ressourcen-Zuteilung in divisional gegliederten Unternehmen	203
HENKE, M.: Mehrstufige Planung einzelner Produkte und Projekte bei Unsicherheit	225
BRAUERS, W.K.: Prospective Planning. A Praktical Application in Defense	246

Lagerhaltung

HOCHSTÄDTER, D.: Die stationäre Behandlung von Mehr-Produkt Lagerhaltungsmodellen	269
---	-----

TAWADROS, M.A.: Sinusoidal Functions for Inventory Control Models	290
Spezielle Aspekte der mathematischen Programmierung	
ECKHARDT, U.: Pseudokomplementärverfahren (Zusammenfassung)	313
STECKHAN, H.: Ein Algorithmus zur Konstruktion von Stromzirkulationen in Netzen	316
JUNGINGER, W.: Über die Lösung des dreidimensionalen Transportproblems	332
GAL, T.: Zur multiparametrischen linearen Programmierung	349
BÜHLER, W.: Zur Lösung eines Zwei-Stufen-Risiko Modells der stochastischen linearen Optimierung	355
Betriebliche Probleme	
MEYHAK, H.: Ein Simulationsmodell zur integrativen Unternehmensplanung	373
HAEHLING von LANZAUER, Ch.: The Effects of the Insurees' Decisions on the Insurers' Profit	399
LIN, C.-Y.: A Decision Theoretic Approach to the Design and Analysis of Industrial Experiments - An Application	422
MENSCH, G.: Zur optimalen Gleitzeitregelung bei stössweisem Arbeitsanfall	437
ORDELHEIDE, D.: Ein Simulationsmodell für den Instandhaltungsbereich	465
Makroökonomische Anwendungen	
BECKER, U.: Zwischenbericht über eine Analyse des Zusammenhanges zwischen Brief- und Fernsprechverkehr im Bereich der Deutschen Bundespost	489
SCHIPS, B.: Stochastische Prozesse und makroökonomische Konjunkturtheorie	501
HENN, R., und O. OPITZ: Dynamische Aspekte der Aktivitätsanalyse	514
BOL, G.: Zur Existenz von Produktionsfunktionen	536
MOESCHLIN, O.: Erweiterungen des Open Expanding Economy Model	552
Stopp-Probleme und Markoff-Modelle	
DIRICKX, Y.M.I.: Linear Programming Algorithms for the Deterministic Discrete Dynamic Programming Problem	563
GOLDSTEIN, B.H.: Über Stop-Probleme bei diskreten Markoff-Ketten	579
UEBE, G.: Classifying the States of a Finite Markov Chain	587

EMRICH, O. : Optimales Stoppen von endlichen Markoff-Ketten	604
Prognose- und Schätzverfahren	
DATHE, H. M. : Die Bedarfsanalyse als Hilfsmittel für technisch- wirtschaftliche Planungen	621
SEILER, K. : Marginal Utility in the Economization of Power Series	658
STIER, W. : Spektralanalytische Untersuchungen von Aktienkurs- entwicklungen	665
Investitionsentscheidungen	
SUTTON, A. M. : The Use of Simulation in the Evaluation of Alternative Designs and the Forecasting of Revenue for High Investment Transport Service Facilities	683
Probleme der Fertigung	
KOSTEN, L. : Heuristische Methoden zur Arbeitsangleichung bei Fließbändern	701
SIERENBERG, R. W. : An Algorithm for the Line Balancing Problem	722
Ganzzahlige Programmierung	
MÜLLER-MERBACH, H. : Modifikationen von Cutting-Plane-Methoden der ganzzahligen Optimierung	747
Autorenverzeichnis	789

Einführung

Unternehmensforschung in Wissenschaft, Wirtschaft und Politik

von H. Künzi, Zürich

1. Einleitung

Die Unternehmensforschung, auch Operations Research genannt, hat ihre Wurzeln weitgehend in der angewandten Mathematik und in der Statistik. Man geht nicht fehl, wenn man die Theorie des neuen Forschungszweiges als Teilgebiet der angewandten Mathematik bezeichnet. Aus dieser Ueberlegung heraus kann man die Unternehmensforschung auch als diejenige Richtung der angewandten Mathematik betrachten, die sich dem Anwendungsgebiet der Wirtschaft und der Politik zuwendet. Sicher gehen theoretische und praktische Ansätze aus dem Gebiete der Unternehmensforschung schon sehr weit zurück; aber das, was wir heute als eigentliche Unternehmensforschung bezeichnen, weist auf eine kaum dreissigjährige Geschichte hin, deren Geburtsort eindeutig in den Vereinigten Staaten zu suchen ist. In Europa hat man sich dann rasch für die wertvollen Gedankengänge interessiert, vor allem die Engländer und Franzosen. In den ersten Jahren dieser Entwicklungsgeschichte beschäftigte man sich vorwiegend mit theoretischen Betrachtungen. Im Vordergrund stand vor allem die Theorie der linearen und nichtlinearen Optimierung. Rasch entstand auch die Spieltheorie und die umfassende Entscheidungstheorie. Das machtvolle Einsetzen des Computers und der Computerwissenschaften gaben der Unternehmensforschung gewaltige Impulse, so zum Beispiel in der Simulationstheorie. Was Deutschland und die Schweiz anbetrifft, muss heute festgestellt werden, dass dieser neue Zweig der angewandten Mathematik relativ spät Fuss fassen konnte. Man beschränkte sich sporadisch auf die Uebernahme einiger amerikanischer Modelle und fand dann sehr oft, dass sich diese nicht ohne weiteres für unsere Verhältnisse eigneten.

Die eigentliche Geburtsstunde in Deutschland wie in der Schweiz kann im Jahre 1961 fixiert werden, als nämlich in beiden Ländern entsprechende Fachverbände ins Leben gerufen wurden.

In Deutschland war es die Deutsche Gesellschaft für Unternehmensforschung (DGU) und auf der südlichen Seite des Rheines die Schweizerische Vereinigung für Operations Research oder SVOR. Die beiden Gesellschaften, die somit auf eine zehnjährige Tätigkeit zurückblicken können, hatten zum Ziele, die Unternehmensforschung theoretisch und praktisch zu fördern. Es muss als Glücksfall bezeichnet werden, dass in beiden Gesellschaften massgebende Persönlichkeiten aus der Wissenschaft und der Wirtschaft aktiv am Aufbau mitarbeiteten. Obschon Skeptiker und Kritiker in beiden Ländern nicht fehlten, entwickelte sich das neue Forschungs- und Anwendungsgebiet rasch. Es entstanden Lehrstühle und Institute, und die fortschrittlichen Wirtschaftsunternehmen sowie politische Instanzen suchten Kontakte und Anknüpfungspunkte im noch bescheidenen Garten der Unternehmensforschung.

Heute darf mit berechtigtem Stolz festgestellt werden, dass sich aus den zarten Pflänzchen "Unternehmensforschung", die vor zehn Jahren in Deutschland und in der Schweiz gepflanzt wurden, ansehnliche Bäume entwickelt haben, die nicht mehr aus dem Landschaftsbild wegzudenken sind. Diese Bäume weisen jetzt schon zahlreiche Aeste und Verzweigungen auf; trotzdem benötigen sie aber auch in Zukunft noch weitere Pflege. Man muss sie vor allem stärken, damit sie den verschiedenen Sturmwehen, die nicht selten im politischen und wirtschaftlichen Landschaftsgarten ihr Unwesen treiben, standhalten können.

2. Operations Research in Wirtschaft und Politik

Nach zehn Jahren intensiver Arbeit in Theorie und Praxis ist festzuhalten, dass in beiden Ländern, wenn auch nicht in allen Sparten, so aber doch in einigen wichtigen Teilgebieten, der Anschluss an die weltweite Entwicklung gefunden werden konnte. Das trifft vor allem in den zentralen Gebieten der Optimierung, der Entscheidungstheorie und teilweise beim Gebiet der Simulation zu. Natürlich verfügen wir noch bei weitem nicht über den Forschungsstand wie das z. B. in den USA der Fall ist; aber an internationalen Kongressen ist die Präsenz aus den beiden Ländern jeweils beachtlich. Wirtschaftsbetriebe verfügen teilweise schon über eigene Operations Research-Gruppen, und verschiedene Unternehmungen lassen sich durch private oder Universitätsinstitute beraten. Die frühere Skepsis ist weitgehend zusammengeschrumpft, und praktisch alle fortschrittlichen Betriebe, die sich mit entsprechenden Problemen auseinanderzusetzen haben, erkennen den Wert der neuzeitlichen Methoden. Etwas zaghafter ist die Politik oder die öffentliche Verwaltung in Richtung des Einsatzes der Unternehmensforschung vorgegangen. Neben Aemtern, die sich gerne beraten lassen, gibt es noch immer recht konservative Stellen, die einer wissenschaftlichen Durchleuchtung politischer Aufgaben ablehnend gegenüberstehen. Wir alle wissen aber, dass zwischen einem umfangreichen wirtschaftlichen Unternehmen und einer grossen öffentlichen Verwaltung weitgehende Parallelen bestehen. Je länger je mehr wird die öffentliche Verwaltung zum eigentlichen Unternehmen, ja zum Grossunternehmen, das nur noch funktionieren kann, wenn es von einem modernen Management und von einer neuzeitlichen Betriebsführung gelenkt wird. Vielleicht bedarf es an solchen Stellen noch in vermehrtem Masse der Orientierung und der Information.

Unternehmensforschung gehört nicht nur in die wirtschaftliche Unternehmung, um diese zum Blühen zu bringen, sondern auch in die öffentliche Verwaltung, um diese zum besseren Funktionieren zu bringen.

3. Ausbildungsprobleme

Wenn heute noch nicht überall von einem durchschlagenden Erfolg der Unternehmensforschung, vor allem im wirtschaftlichen Sektor, gesprochen werden kann, so ist dies sicher in hohem Masse darauf zurückzuführen, dass leider in Deutschland und in der Schweiz noch recht wenig erstklassige Spezialisten vorhanden sind. Das bezieht sich sowohl auf die Wissenschaftler wie auf die entsprechenden Praktiker. Zahlreiche Projekte, die sich hervorragend für Operations Research-Studien geeignet hätten, konnten aus diesem Grunde nicht behandelt werden. Natürlich brauchen wir in der Unternehmensforschung nicht ausschliesslich Theoretiker, die die entsprechende mathematische oder statistische Seite beherrschen, sondern man ist in hohem Masse auf solche Spezialisten angewiesen, die einerseits die Theorie kennen, aber daneben in der Lage sind, auch das praktische wirtschaftliche oder volkswirtschaftliche Problem zu überblicken, um darauf den theoretischen Apparat anzuwenden. Der eigentliche Spezialist in der Unternehmensforschung muss daher vielseitig sein, theoretisch gründlich ausgebildet und in den verschiedenen praktischen Sparten bewandert. Leider fehlen uns in Deutschland und in der Schweiz noch weitgehend typische Ausbildungsstätten der Unternehmensforschung. Der Einzelne ist auf sich allein angewiesen; er muss sein Studium aus mehreren Fakultäten selber zusammentragen, was sicher keinen allzugrossen Anreiz für diese Fachrichtung auszuüben vermag. Meines Erachtens sollte in der Neugestaltung der Lehrpläne an unseren Universitäten dieser offensichtlichen Lücke Rechnung getragen werden. Es muss untersucht werden, ob ein Studium der Unternehmensforschung, das sich auf Mathematik, Wirtschaftswissenschaft und Technik abstützt, sich als sinnvoll erweist, oder ob die Ausbildung - ich persönlich weise eher in diese Richtung - in Form von "post graduate" Studien anzustreben sei. So oder so benötigen

wir sicher in Deutschland wie in der Schweiz wissenschaftliche Zentren, die im Gebiete der Unternehmensforschung eng miteinander verbunden sind.

4. Erreichtes in der Unternehmensforschung

Es würde den Rahmen dieses Referates bei weitem sprengen, wenn ich das bisher Erreichte im Bereiche der Unternehmensforschung in Deutschland und der Schweiz auch nur kurz skizzieren wollte. Vielmehr liegt mir daran, einige Schwerpunkte aufzuzeichnen, mit denen ich durch die Aktivität meines früheren Institutes für Operations Research und elektronische Datenverarbeitung der Universität Zürich konfrontiert wurde. Ich beschränke mich auch, und dies wiederum im Zusammenhang mit meiner jetzigen Tätigkeit im öffentlichen Dienst, auf vorwiegend staatliche Probleme.

A) Volkswirtschaftliche Probleme

In Zusammenarbeit mit dem eidgenössischen Volkswirtschaftsdepartement haben wir in den letzten Jahren die modernsten mathematischen und technischen Hilfsmittel eingesetzt, um zentrale volkswirtschaftliche Projekte, vorwiegend auf dem Gebiet der Kriegsvorsorge, lösen zu helfen; an erster Stelle erwähne ich den sogenannten schweizerischen Anbauplan. Seine Aufgabe besteht darin, die Bebauung des knappen schweizerischen Bodens so zu gestalten, dass unsere Bevölkerung in Notzeiten aus der eigenen Produktion ernährt werden kann. In Analogie zum bekannten "Plan Wahlen" aus dem zweiten Weltkrieg wurde ein grosses Modell mit mehreren hundert Relationen und Variablen aufgestellt. Dieses umfangreiche Modell, das sich exakter mathematischer Methoden bedient, ist in der Lage, das mathematische Optimum mit Sicherheit anzugeben. Die Berechnung kann

nur auf einer sehr leistungsfähigen elektronischen Rechenanlage erfolgen. Die Rechenzeit beläuft sich auf ungefähr eine Stunde. Dabei muss natürlich erwähnt werden, dass die Ausfertigung des Computerprogramms eine umfangreiche Arbeit darstellte. Doch dieses Programm kann, wenn es einmal vorhanden ist, immer wieder benutzt werden.

Neben diesem Anbauplan werden im Rahmen der neugeschaffenen Sektion KOR (Kriegswirtschaftliches Operations Research) verschiedene weitere Projekte studiert, die hier nur kurz erwähnt werden sollen:

- a) Vorbereitung einer Lebensmittelrationierung im Kriegsfall unter Einsatz von Computern
- b) Probleme der Mehl- und Brotversorgung. Dabei gilt es zu ermitteln, auf welche Weise die Brotversorgung der Schweiz im Kriegsfall kostenoptimal gesichert werden kann. Fragen treten auf in der Form: Welche Ortsgetreidestelle hat welche Mühle zu beliefern, und zu welchem Verbraucherzentrum wird das Mehl dann weitergeleitet, so dass der Mehlbedarf gedeckt wird und die Kosten minimal werden.
- c) Zu unseren interessanten Problemen gehört der Operations Research-Teil des zur Zeit im Auftrage stehenden Agrargutachtens des Bundesrates über die langfristige Struktur der schweizerischen Landwirtschaft. Hier steht die optimale Produktionsstruktur für zwei Varianten im Vordergrund, nämlich
 - 1) bei einem allfälligen EWG-Anschluss unter spezieller Berücksichtigung der EWG-Preise,
 - 2) bei keinem EWG-Anschluss.

Auch bei dieser Aufgabe ist die Kriegsvorsorge zu berücksichtigen, im Zusammenhang mit der minimalen Anbaubereitschaft.

- d) Die Nachfragefunktionen für die ca. 20 wichtigsten Agrarprodukte wurden mittels ökonomischer Methoden geschätzt. Die Kenntnis solcher Nachfragefunktionen der wichtigsten Agrarprodukte kann der schweizerischen Agrarpolitik bestimmte wichtige Dienste leisten.

- e) Besonderes Interesse möchten wir der Aufstellung volkswirtschaftlicher Gesamtmodelle widmen. Solche Modelle haben sich vor allem mit dem Problem der allgemeinen Interdependenz, d.h. mit der Tatsache, dass alle wirtschaftlichen Vorgänge miteinander verkettet sind, auseinanderzusetzen.

B) Militärische Probleme

a) Evaluation von Kampfflugzeugen

Bei diesen Operations Research-Studien geht es im Wesentlichen darum, aus den dem Eidgenössischen Militärdepartement angebotenen Flugzeugtypen denjenigen Typ zu ermitteln, der die verlangten militärischen Wirkungen mit den relativ geringsten Kosten erzielt. Zur Untersuchung eines sehr wichtigen Teilproblems, nämlich der Ueberlebenswahrscheinlichkeit eines Kampfflugzeuges im Kriegseinsatz, wurde ein Operations Research-Modell zur Simulation von Luftkämpfen auf dem Elektronenrechner entwickelt. Für dieses Modell wurden die Einflussfaktoren auf die Bewegung, die Sichtung und auf den Luftkampf zweier Flugzeuge oder Flugverbände untersucht. Durch das möglichst realistische Nachbilden (Simulieren) von Luftkämpfen mit dem Computer soll insbesondere die Bedeutung wichtiger technischer Eigenschaften, wie zum Beispiel das Beschleunigungsvermögen, für die Ueberlebenswahrscheinlichkeit eines Kampfflugzeuges abgeschätzt werden. Je grösser seine Lebenserwartung im Kriegseinsatz ist, umso mehr Einsätze können mit demselben Flugzeug geflogen werden, bevor es durch Abschuss zerstört wird. Um zuverlässige Aussagen über die Ueberlebenswahrscheinlichkeiten erhalten zu können, mussten auf einem Computer gegen 100 000 Luftkämpfe simuliert werden.

b) Weitere militärische Operations Research-Aufgaben

In Kürze sei noch auf einige Studien von militärischen Operations Research-Arbeiten stichwortartig hingewiesen, die in

den letzten Jahren in der Schweiz behandelt wurden: Probleme der Munitionslagerung - Simulation von Bombardierungen eines Flugplatzes - optimaler Erneuerungszyklus bei Motorfahrzeugen - Transportmodelle - Pertstudien - Panzerabwehrsimulation - Modelle zur Tieffliegererfassung u. a. m.

C) Verkehrsplanungen

Die zukünftige Bewältigung des Strassenverkehrs stellt die kommunalen, kantonalen und eidgenössischen Behörden vor mannigfaltige Aufgaben, die nur in Zusammenarbeit von Vertretern aus verschiedenen Fachgebieten mit Aussicht auf Erfolg in Angriff genommen werden können. Für solche umfangreiche Arbeiten stellt das Operations Research wiederum geeignete Verfahren und Modellstrukturen zur Bearbeitung zur Verfügung. Ebenso erweisen sich die modernen Hochleistungs-Rechenautomaten als ein äusserst wertvolles Hilfsmittel, welches zur Bearbeitung eines sehr grossen Aufgabenspektrums verwendet werden kann.

In den grossen Städten erfordert der öffentliche Verkehr eine zunehmend gründlichere Verkehrsplanung. Es ist zum Beispiel bei der Entwicklung einer Verkehrssteuerung das Ziel, den Verkehrsablauf durch geeignete Massnahmen nach bestimmten Gesichtspunkten zu optimieren (minimale Wartezeit, grösstes Verkehrsvolumen, etc.). Analytische Verfahren eignen sich nicht immer zur Behandlung solch komplexer Probleme; deshalb benützt man oft grosse Elektronenrechner und simuliert mit Operations Research-Methoden den zu untersuchenden Prozess. Beim Aufstellen von Simulationsmodellen zeigen sich zudem oft weitere, bis dahin noch nicht erfasste Zusammenhänge. Aus den Simulationsergebnissen können

Schlüsse auf eine günstige Fahrplangestaltung der Strassenbahnen gezogen werden; ebenso vermitteln die Simulationsergebnisse Angaben über die Auswirkungen der untersuchten Ampelsteuerstrategien auf die übrigen Verkehrsteilnehmer, sowie Hinweise auf allfällige notwendige bauliche Veränderungen.

D) Bauplanung

Die Investitionen im Bauwesen sind in den letzten Jahren enorm angestiegen. Durch eine sorgfältige Auswahl von Standorten sowie durch zweckmässige Raumdimensionierungen und Raum-anordnungen kann der Nutzwert von Gebäuden wesentlich erhöht werden. Bezüglich der Baukosten ist zu bemerken, dass durch die Wahl geeigneter Bauverfahren, durch die Bestimmung optimaler Seriengrössen von Bauteilen und durch die Wahl einer optimalen Baugegeschwindigkeit wesentliche Kosteneinsparungen erreicht werden können.

Bei den Berechnungen wird ein gegebenes Zahlenmaterial unter vielen Gesichtspunkten ausgewertet und mittels Verfahren des Operations Research optimiert. Die Auswahl der Rechenoperationen erreicht in der Regel einen Umfang, der nur durch den Einsatz elektronischer Grossrechenanlagen in wirtschaftlicher Weise bewältigt werden kann.

Es lag mir daran, an diesen wenigen Beispielen den Einsatz der Unternehmensforschung in der Schweiz zu erläutern.

Oft mutet es sonderbar an, dass diese wirkungsvollen Methoden nicht schon früher verwendet wurden. Ich glaube, dass unsere grossen Mathematiker des letzten Jahrhunderts ohne weiteres in den

Lage gewesen wären, den erforderlichen theoretischen Apparat zu liefern. Der Grund dieses späten Einsatzes des Operations Research lässt sich aber ganz einfach erläutern; Ohne Computereinsatz lohnt es sich nicht, Operations Research zu betreiben. Es lässt sich eindeutig feststellen, dass die Geburtsstunde des Operations Research ziemlich genau mit der Geburtsstunde des Computers zusammenfällt.

Es soll unser Bestreben sein, diese beiden Komponenten, nämlich das Operations Research und den Computer, am richtigen Ort einzusetzen und auf die richtige Weise zu fördern. Die obigen Beispiele, die sich durch zahlreiche andere noch erweitern liessen, haben uns eindrücklich vor Augen geführt, wie nützlich diese beiden Komponenten sind.

5. Der Ausbau der Unternehmensforschung auf nationaler Ebene

In den vorangegangenen Abschnitten glaube ich deutlich gezeigt zu haben, wie sehr weite Kreise heutzutage auf die Unternehmensforschung angewiesen sind. Die Konsequenz daraus lautet: Förderung dieses Zweiges sowohl in wissenschaftlicher wie auch in praktischer Richtung auf breiter Ebene.

Die beiden Vereinigungen tun das ihrige dazu. Neben der DGU und der SVOR bemühen sich noch weitere Institutionen in verdankenswerter Weise für einen gezielten Ausbau der Unternehmensforschung. In hohem Masse sehe ich darin aber eine staatliche Aufgabe. Was dem Staate direkt und indirekt nützt, hat er entsprechend zu fördern.

Unter Mitwirkung von Kollege Professor W. Krelle habe ich ein Programm ausgearbeitet, das der staatlichen Förderung der Wirtschaftswissenschaften im allgemeinen und der Unternehmensforschung im speziellen in Deutschland und der Schweiz Rechnung trägt.

Es kann wie folgt zusammengefasst werden:

1. Sowohl in Deutschland wie auch in der Schweiz sollten je ein oder zwei Zentren für Unternehmensforschung an Universitäten oder Technischen Hochschulen errichtet werden. An diesen Zentren sollten diejenigen Unternehmensforscher vereinigt werden, die Theorie und Praxis von hoher Warte aus betreiben. Die Berufung an diese Zentren sollte durch ein unabhängiges Gremium, das auch die hervorragendsten ausländischen Fachvertreter umfasst, auf Zeit erfolgen (ca. fünf Jahre). Nach dieser Zeit tritt der Berufene wieder in seine alte Stellung zurück.
Eine genügende Zahl von Stellen für Assistenten und Schreibkräfte müssen für das Zentrum zur Verfügung stehen.
2. Es sollten alle möglichen Massnahmen ergriffen werden, um eine Verbindung von reiner Forschung und Praxis herzustellen. Hierzu wird vorgeschlagen:
 - a) Regelmässige Abordnung von leitenden Beamten in der Verwaltung und Persönlichkeiten in Führungspositionen der Wirtschaft zu Kursen an der Universität, in denen sie mit dem Fortgang der Wissenschaft vertraut gemacht werden und ihrerseits die Forschung auf dringende Probleme der Praxis hinweisen.
 - b) Zeitweise Abordnung von Wissenschaftern in die Verwaltung, um bestimmte Planungsaufgaben zu leiten, für die sie ihrer Forschungsrichtung nach kompetent sind.
3. Der Staat sollte die Unternehmensforschung in erheblichem Masse fördern. Besonderes Gewicht ist auf die Grundlagenforschung zu legen. Die angewandte Forschung könnte durch direkte Regierungsaufträge für bestimmte Projekte, ähnlich wie in den USA, gefördert werden.

Die wirtschaftliche Zukunft beider Länder hängt vom Stand der Wissenschaft und der Technik ab. Die wissenschaftliche Vorherrschaft der USA ist weitgehend durch die grössere Förderung der Forschung durch die Regierung bedingt. Unsere Regierungen sollten hier entsprechend gleichziehen, damit wir nicht weiter zurückfallen.

4. Sowohl in Deutschland wie in der Schweiz sollte die Studienreform für das Gebiet der Unternehmensforschung vorangetrieben werden.
5. Die Zentren in Deutschland und in der Schweiz sollten eng zusammenarbeiten, um dadurch eine noch grössere Ausstrahlungskraft zu gewinnen.

Wiederum erachte ich die DGU und die SVOR als Institutionen, die wesentlich zur Realisation dieses Programms beitragen können, das sicher dafür Gewähr bieten würde, dass die Unternehmensforschung in unseren beiden Ländern die ihr zustehende Bedeutung erlangen wird.

Beziehungen zwischen Mathematik und Wirklichkeit in mathematischen Modellen von empirischen Strukturen

von A. Jaeger, Bochum

1. EINLEITUNG

Der amerikanische Mathematiker G. B. DANTZIG bemerkte einmal in einem Interview mit einem Journalisten, man könne mit einem Modell aus dem von ihm entwickelten Gebiet der linearen Programmierung streng beweisen, daß von allen möglichen Arten des Zusammenlebens der Geschlechter in einer sozialen Gruppe (Monogamie, Bigamie, Polygamie usw.) die Monogamie optimal sei. Darauf soll der Journalist nur kopfschüttelnd mit dem Kommentar reagiert haben: Dann müsse DANTZIG sicherlich mit der falschen Art von Modellen gearbeitet haben ([2], S. 367).

Diese köstliche Geschichte ereignete sich schon vor mehr als fünfzehn Jahren. Heutzutage sind weitaus vielfältigere Mißverständnisse bei Verwendung der Vokabel *Modell* denkbar, da sich diese zu einem allseits beliebten Modewort mit wissenschaftlichem Nimbus entwickelt hat. Die umgangssprachliche Abgegriffenheit dieser Vokabel kann nun leicht zu dem Fehlschluß führen, daß sich auch der Begriff des abstrakten Modells im Rahmen einer Wissenschaft eigentlich nur in eine Metapher oder in ein vages Schlagwort auflöst. Mit Hilfe weniger elementarer mathematischer Grundbegriffe läßt sich jedoch der wissenschaftliche Modellbegriff präzisieren. Der wesentliche Hinderungsgrund für das Verstehen der exakten Definition des mathematischen Modelles besteht für den Nicht-Spezialisten lediglich da-

rin, daß das Stichwort *mathematisch* bei ihm eine sofortige abweisende Reaktion auslösen kann. Denn in der Schule mag das Fach, welches unter der Bezeichnung *Mathematik* gelehrt wird, meist nur Schrecken und ein Gefühl der Ohnmacht vermittelt haben, und viele sogenannte "Gebildete" wollen mathematische Überlegungen gar nicht anstellen, ja geben sogar ein Fehlen von mathematischen Kenntnissen ohne Scham und fast mit Stolz zu, während sie wohl gleichzeitig Lücken auf schöngeistigen Gebieten sorgfältig zu verbergen suchen. Es kann hier nicht der Ort sein, im einzelnen zu analysieren, worauf diese bedauerliche Entwicklung gerade in der heutigen, so sehr durch die exakten Naturwissenschaften geprägten Zeit zurückzuführen ist; aber vielleicht muß man ungenügender Relevanz des Mathematikunterrichtes zur realen Welt einen erheblichen Anteil der Verantwortung zuschreiben. Dabei liegt, wie in diesem Referat behutsam entwickelt werden soll, in dem Verstehen des Zusammenhanges zwischen Wirklichkeit und Mathematik gerade der Schlüssel zum Verständnis des Modellbegriffs der Wissenschaft.

2. ABSTRAKTIONEN UND KLASSEN

Wie die Entwicklung der menschlichen Sprache, mit ihren akustischen und später auch optischen Symbolen, durch eine Folge von Verfeinerungen von Abstraktionen aus Gegebenheiten der realen Welt erfolgt, so geschieht dies auch mit der Entwicklung jeder Wissenschaft, einschließlich der Mathematik. Abstraktion bedeutet hierbei das Weglassen von Eigenschaften, die Trennungsmerkmale beinhalten, so daß

die dann verbleibenden Eigenschaften Gemeinsamkeiten signalisieren. Unter Verwendung des Begriffes der *M e n g e*, dem heutzutage ja schon immer mehr die Schulkinder (und deren Eltern) ausgesetzt sind, wenn sie die sogenannte *Neue Mathematik* lernen müssen, läßt sich jeder Abstraktionsvorgang dahingehend interpretieren, daß jedem Element einer Menge von Gegebenheiten der realen Welt oder von Begriffen genau ein Element einer neuen Menge von abstrakteren Begriffen zugeordnet wird. Durch diese *A b b i l d u n g*, wie der Mathematiker gewöhnlich eine derartige Zuordnung nennt, wird im allgemeinen verschiedenen Elementen der Ausgangsmenge ein und dasselbe *B i l d* der neuen Menge zugeteilt. So kann man etwa, ausgehend von der Menge der verschiedenen Baumarten, so z.B. von den Begriffen Tanne, Fichte, Kiefer, zu dem Oberbegriff Nadelbaum, aber z.B. von den Begriffen Eiche, Buche, Birke, zu dem Oberbegriff Laubbaum und danach von den Begriffen Nadelbaum und Laubbaum zu dem Oberbegriff Baum übergehen (siehe Fig. 1).

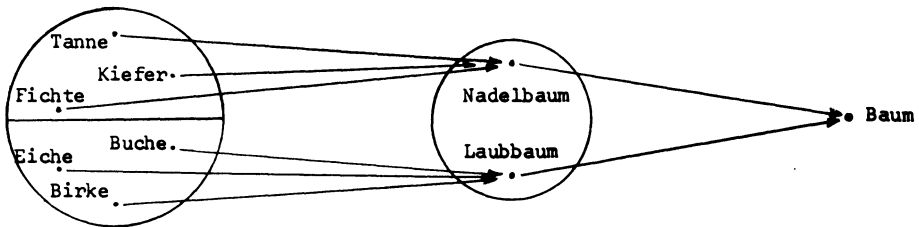


Fig. 1 . Abstraktionsschritte in der Botanik

Jeder solche Abstraktionsschritt führt zu einer Vergrößerung der Beschreibung und Auffassung der betrachteten Objekte, z.B. Gegebenheiten der realen Welt, bei welcher das ursprüngliche Gleichsein dieser Objekte zu einem Ähnlichsein - in bezug auf die übriggebliebenen Gemeinsamkeiten, also hier im Beispiel die Gattungseigenschaften von Bäumen - vergrößert wird. Die Elemente der Ausgangsmenge

werden hierdurch in **K l a s s e n** aufgespalten, wobei jede Klasse genau aus all denjenigen Elementen besteht, denen dasselbe Bild, also dasselbe abstraktere Objekt, zugeordnet ist.

3. AUSSAGEFORMEN UND RELATIONEN

Der menschliche Geist konstruiert nun aber Abstraktionen nicht - oder wenigstens vorwiegend nicht nur - um ihrer selbst willen, er bedient sich ihrer auch, um Eigenschaften, Beziehungen, Gliederungen, Gesetzmäßigkeiten, Zusammenhänge, kurz Organisationsmuster oder -merkmale, zu finden oder zu präzisieren. Eine solche Beziehung trat soeben schon auf, die Beziehung der Zugehörigkeit von zwei Elementen einer Menge zu einer Klasse im obigen Sinne. Derartige Beziehungen lassen sich mit Hilfe der Sprache (der Umgangssprache oder einer wissenschaftlichen Fachsprache mit ihren Fachsymbolen, einschließlich der mathematischen Sprache, je nach den zur Betrachtung stehenden Objekten) folgendermaßen beschreiben: Man stelle eine Behauptung auf, die grammatikalisch wie ein Satz aussieht, aber etwa in derselben Weise variable Größen enthält wie z.B. die sogenannten "Unbekannten" x, y, \dots bei mathematischen Gleichungen, oder wie die "großen Unbekannten" X, Y bei den Steckbriefen von Verbrechern. Genauer gesagt, derartige Behauptungen sollen ein Symbol oder mehrere Symbole als grammatikalischen Satzteil oder als Satzteile enthalten, von denen jedes jeweils stellvertretend für ein beliebiges Element einer gleichzeitig spezifizierten Menge steht. Unabhängig davon, für welchen Wahrheitsbegriff man sich entschieden hat (ob, grob ausgedrückt: wahr z.B. logisch beweisbar oder experimentell verifizierbar bedeuten soll), wird die Frage nach der Wahrheit einer solchen mit Variablen behafteten Aussage, die **A u s s a g e f o r m** ([5], S. 40) oder auch **S a t z** -

form, Aussagefunktion oder Satz-funktion genannt wird, im allgemeinen erst dann einen Sinn haben, wenn man die Variablen durch fest herausgegriffene Elemente der zugrundegelegten Mengen ersetzt, "durch Konstante spezialisiert" (siehe [4], S. 12). Steht z.B. x stellvertretend für eine Zahl (genauer: für ein Element der Menge aller Zahlen), so ist die Aussageform " x ungleich 3" falsch für die Spezialisierung $x = 3$ (d.h., wenn die Variable x durch die Konstante 3 ersetzt wird), aber richtig für alle anderen Spezialisierungen. Oder steht X stellvertretend für eine Person (genauer: für ein Element aus der Menge aller Menschen) und Y für eine Beschäftigung (genauer: für ein Element aus der Menge aller Beschäftigungen), so ist " X übt zur Zeit die Beschäftigung Y aus" eine Aussageform mit zwei Variablen. Beispiele anderer Aussageformen, zu denen sich jeweilige Bezugsmengen leicht konstruieren lassen:

" $x + y = z$ ",

" X hat an der Universität U am Tage T die Prüfung P bestanden",

" X ist mit Y verheiratet",

" X ist Vater von Y ",

"Der Beobachtungswert x ist gleich dem Beobachtungswert y gemessen worden" (Messgleichheit, d.h. Gleichheit im Rahmen der Meßgenauigkeit).

"Der Computer hat ausgerechnet, daß x denselben Wert hat wie y " (Rechengleichheit, d.h. Gleichheit im Rahmen der Rechengenauigkeit).

" X ist Oberbegriff von Y ".

Durch jede derartige Aussageform mit einer Bezugsmenge oder bei mehreren Variablen auch mit mehreren (nicht notwendigerweise verschiedenen) Bezugsmengen ist nunmehr eine Beziehung zwischen den Elementen dieser Mengen gewonnen worden, die *n-stellig*

Relation in oder auf der Bezugsmenge bzw. zwischen den Bezugsmengen genannt wird, wenn die Anzahl der darin auftretenden verschiedenen Variablen ist ($[1]$, S. 256, $[4]$, S. 101 u. S. 109). Für den Spezialfall einer einstelligen Relation ist auch die Bezeichnungsweise Eigenschaft üblich. Man sagt, daß (bei $n = 1$) ein Element s aus der Bezugsmenge bzw. (bei $n > 1$) eine Folge (z.B. ein Paar, ein Tripel, ein Quadrupel usw.) s_1, s_2, \dots, s_n von Elementen aus den jeweiligen Bezugsmengen die Relation erfüllt oder zulässig für die Relation ist, wenn bei der Spezialisierung: x durch s ersetzt (" $x = s$ "), bzw. den Spezialisierungen: x_1 durch s_1 ersetzt, x_2 durch s_2 ersetzt, ..., x_n durch s_n ersetzt (" $x_1 = s_1, x_2 = s_2, \dots, x_n = s_n$ ") die Aussageform in eine Aussage umgewandelt wird, welche wahr ist. Wegen dieses Zusammenhanges kann man eine Aussageform auch eine Relationsbehauptung für die Bezugsmenge(n) nennen ($[4]$, S. 146). (Genauer gesagt: der Mathematiker *definiert* die Relation einfach als Zusammenfassung derjenigen Elemente, bzw. Elementepaare bzw. Elementefolgen, für welche die zugrunde liegende Behauptung wahr ist, also als die Menge der zulässigen Elemente, bzw. Elementepaare bzw. Elementefolgen.)

4. STRUKTUREN UND STRUKTURKOMPLEXE

Betrachtet man eine Menge M zusammen mit einer Relation R oder mit mehreren Relationen R_1, R_2, \dots , die durch Behauptungen mit Variablen aus dieser Menge erklärt sind, so spricht man von einer strukturierten Menge, kürzer von einer Struktur, und schreibt abkürzend (M, R) bzw. $(M; R_1, R_2, \dots)$. (Man findet auch andere Bezeichnungsweisen, wie z.B. Relationengebilde,

r e l a t i o n a l e s S y s t e m , R e l a t i v , S y -
 s t e m , G a n z h e i t , G e s t a l t o d e r i n s p e z i e l l e n
 F ä l l e n G r a p h , G r a p h o i d . Siehe z.B. [1], S. 209
 und 284, [4], S. 119, [5], S. 60) Manchmal muß man sogar mehrere
 Mengen M_1, M_2, \dots , zusammen mit mehreren Relationen, die sich je-
 weils auf einzelne oder alle dieser Mengen beziehen, zugleich be-
 trachten, z.B. wenn man Wertordnungen, Bewertungen (siehe [4], S. 17)
 für Dinge, Phänomene, Ergebnisse, Alternativen usw. einführen will;
 in einem solchen Falle schreibt man natürlich abkürzend (M_1, M_2, \dots ;
 R_1, R_2, \dots). Auch dann soll die Zusammenfassung dieser Mengen und
 der mit ihnen betrachteten Relationen S t r u k t u r o d e r v i e l -
 l e i c h t b e s s e r S t r u k t u r k o m p l e x (wenn man ausdrück-
 lich betonen will, daß mehrere Mengen betrachtet werden) genannt
 werden. Für die zugrundeliegenden Mengen benutzt man Vokabeln wie
 G r u n d m e n g e (n) , B a s i s m e n g e (n) o d e r
 S t r u k t u r t r ä g e r . Das klingt vielleicht bisher alles
 nicht übermäßig mathematisch, denn es wurde bewußt der mathemati-
 sche Charakter dieser Überlegungen selbst auf die Gefahr verbaler
 Unschärfe hin heruntergespielt; aber es sollte mit Nachdruck darauf
 hingewiesen werden, daß die moderne Mathematik nicht nur solche
 Strukturen (in diesem Sinne), die auf "Zahlen und Figuren" aufge-
 baut oder von "Zahlen und Figuren" abgeleitet sind, untersucht, son-
 dern deradezu als die *Theorie abstrakter Strukturen an sich* angese-
 hen werden kann. Vielleicht sollte man sogar die oft emotional - ne-
 gative Vokabel *Mathematik* durch eine neue Wortbildung wie etwa *Struk-*
turik ersetzen.

5. MORPHISMEN VON STRUKTUREN

Häufig beobachtet man nun zwei Strukturen gleichzeitig, vor allem wenn sie eine gewisse Art von *struktureller Ähnlichkeit* aufweisen, die zunächst für den Fall einer einzigen Grundmenge und einer einzigen Relation bei beiden Strukturen vorgestellt werden soll: Es läßt sich eine Abbildung von der zugrundeliegenden Menge M der einen Struktur, *Ausgangs- oder Originalstruktur* genannt, in die zugrundeliegende Menge N der anderen Struktur, *Bildstruktur* genannt, finden, welche die besondere Eigenschaft aufweist, daß aus der Gültigkeit der Relation R der ersten Struktur, der *Ausgangs- oder Originalrelation*, für gewisse Elemente von M automatisch die Gültigkeit der Relation S der zweiten Struktur, der *Bildrelation*, für die *entsprechenden* Bildelemente in N folgt. Dazu ist es allerdings unbedingt erforderlich, daß die zwei zu vergleichenden Relationen dieselbe *Stellenzahl* haben, d.h. daß S auch n -stellig sein muß, wenn R n -stellig ist. Wenn in diesem Sinne die Bildrelation gewissermaßen die Originalrelation *widerspiegelt*, spricht man von einer *kompatiblen Strukturabbildung* oder von einem *(Struktur-) Morphismus* oder *Homomorphismus* von (M,R) in (N,S) (siehe z.B. [1], S. 288, [4], S. 123).

Bedeutet z.B. R die Relation der Gleichheit, definiert durch die Aussageform " x ist gleich y ", in der Menge M , und bedeutet S die Relation der Gleichheit, aber jetzt für die Menge N definiert, so ist ganz offensichtlich *jede beliebige Abbildung* von M in N ein Morphismus von (M,R) in (N,S) . Bedeutet dagegen S die Relation der Ungleichheit für N , definiert durch die Aussageform " x ist nicht gleich y ",

so ist nun plötzlich *keine* Abbildung von (M,R) in (N,S) ein Morphismus.

6. ISOMORPHISMEN VON STRUKTUREN

Bei Abbildungen von Mengen kann man in vielen Fällen durch Änderung der Zuordnungsrichtung wieder zu einer Abbildung, der *i n v e r - s e n* oder *u m g e k e h r t e n* Abbildung, gelangen, nämlich genau dann, wenn in der zweiten Menge N jedes Element ein Bild eines Elementes der ersten Menge, aber niemals Bild mehrerer Elemente der ersten Menge M ist. Bei dieser Umkehrung vertauschen sich dann gewissermaßen die Rollen von Bild und Original. Eine solche spezielle Abbildung nennt der Mathematiker *e i n d e u t i g*, *b i - j e k t i v* oder einfach *u m k e h r b a r*. Typisches Beispiel ist die Partnerzuordnung in der Tanzschule, wenn kein Mauerblümchen übrig bleibt. Ist nun eine Abbildung M in N , die einen Morphismus von (M,R) in (N,S) ergibt, umkehrbar, so ist die Frage interessant, ob auch die umgekehrte Abbildung einen Morphismus bildet, und zwar natürlich dann von (N,S) in (M,R) . In einem solchen speziellen Falle hat man es gewissermaßen mit *struktureller Gleichheit* zu tun und spricht von einem *I s o m o r p h i s m u s* der Ausgangsstruktur *a u f* die Bildstruktur (siehe z.B. [1], S. 290, [4], S. 123, [5], S. 129).

Ein typisches, wohlbekanntes Beispiel für einen Isomorphismus läßt sich mit Hilfe des Oberganges von den positiven Zahlen zu ihren Logarithmen konstruieren; er bildet den Hintergrund für die Benutzung des Rechenschiebers oder der Logarithmentafel. Genauer gesagt, geht es hierbei um folgendes (siehe Fig. 2): Man nehme für M die Menge aller positiven Zahlen, für R die dreistellige Relation, welche durch

die Relationsbehauptung " $x \cdot y = z$ " definiert ist, für N die Menge aller (reellen) Zahlen, für S die dreistellige Relation, welche durch die Relationsbehauptung " $u + v = w$ " definiert ist, sowie diejenige Zuordnung, welche jeder Zahl a ihren Logarithmus: $\log a$ zuordnet. Wenn nun für drei positive Zahlen $a_1 \cdot a_2 = a_3$ gilt, so gilt für ihre Logarithmen: $\log a_1 + \log a_2 = \log a_3$, und umgekehrt. Also ist diese Zuordnung ein Strukturisomorphismus von der Struktur (M, R) auf die Struktur (N, S) . (Dieses Beispiel zeigt übrigens auch gleichzeitig, auf welche Weise sich die elementaren Rechenoperationen der Schule als dreistellige Relationen auffassen lassen.)

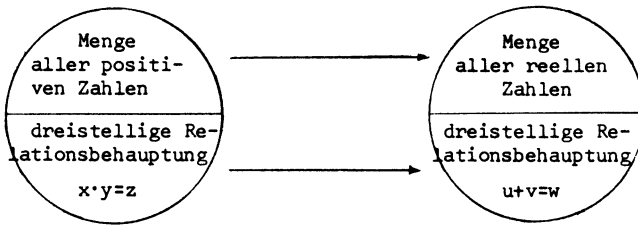


Fig. 2 Das Rechnen mit Logarithmen

7. MORPHISMEN VON STRUKTURKOMPLEXEN

Diese Begriffsbildungen lassen sich nun in naheliegender Weise auf Strukturkomplexe übertragen, wobei die verbale bzw. formelmäßige Beschreibung eigentlich komplizierter ist als der zugrundeliegende Gedanke. Um die Unübersichtlichkeit nicht allzu groß werden zu lassen, soll allerdings zunächst angenommen werden, daß die zu betrachtenden Strukturkomplexe jeweils nur eine zugrundeliegende Menge ha-

ben. Man gehe aus von einem Strukturkomplex $(A; P_1, P_2, \dots, P_p)$, welcher sich aus der Menge A und einer Anzahl von Relationen P_1, \dots, P_p in dieser Menge zusammensetzt, und bringe diesen Strukturkomplex mit einem zweiten (i.a. anderen) Strukturkomplex $(B; Q_1, \dots, Q_q)$ in Verbindung, welcher sich seinerseits aus der Menge B und den Relationen Q_1, \dots, Q_q in dieser Menge zusammensetzt. Dabei braucht keinesfalls verlangt zu werden, daß die Anzahl der Relationen des ersten Strukturkomplexes gleich der Anzahl der Relationen des zweiten Strukturkomplexes ist, also unbedingt $p = q$ erfüllt sein muß. Es soll nun von einer Abbildung des Ausgangskomplexes in den Bildkomplex gesprochen werden, wenn erstens eine Abbildung der Menge A in die Menge B festgelegt ist und zweitens jeder Relation der Ausgangsstruktur eine Relation der Bildstruktur mit *gleicher* Stelenzahl zugeordnet ist. Dabei ist keineswegs vorausgesetzt, daß zwei verschiedenen Ordinalrelationen unbedingt zwei verschiedene Bildrelationen zugeordnet sein müssen. Nun besteht wiederum der Wunsch, die verschiedenen Zuordnungen so vorzunehmen, daß sich die Bildstruktur irgendwie *ähnlich* wie die Originalstruktur verhält; genauer gesagt, es ist nun das Folgende erwünscht: Erfüllen Originalelemente eine Originalrelation, so sollten die entsprechenden Bildelemente die entsprechende Bildrelation erfüllen. Trifft das allgemein zu, so wollen wir *in wichtiger Verallgemeinerung der üblichen Definition* weiterhin von einem Morphismus oder von Kompatibilität sprechen.

BEISPIEL: A sei eine Menge verschiedener (physikalischer, chemischer, ökonomischer, soziologischer oder anderer) Zustände. Durch Meß- oder Beobachtungsanordnungen werden numerische Daten bestimmt, wobei angenommen werden möge, daß verschiedene Zustände verschiedene Meßdaten ergeben. Es seien die Relation P_1 durch die Relationsbehauptung "x und y ergeben denselben Meßwert", die Relation P_2 durch die Re-

lationsbehauptung "x und y sind gleich" definiert. Die Menge B bestehe aus den obigen Meßdaten, Q sei die Relation der Gleichheit in B. Unter diesen Voraussetzungen ergibt sich ein Strukturmorphismus von $(A; P_1, P_2)$ in $(B; Q)$ in dem von uns definierten Sinne, wenn jedem Zustand aus A sein Meßwert in B und *beiden* Relationen P_1, P_2 die Relation Q zugeordnet wird (siehe Fig. 3).

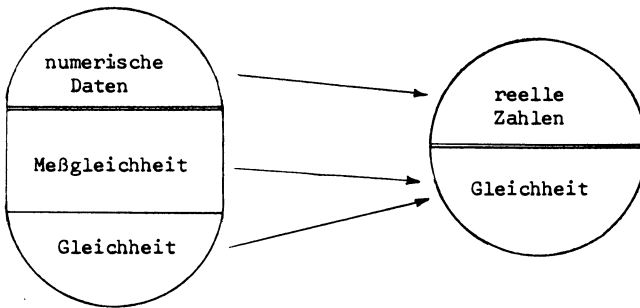


Fig. 3 Ein Strukturmorphismus

Das folgende Beispiel ist sehr ähnlich. A sei die Menge aller reellen Zahlen. Ein Computer soll Rechenoperationen ausführen, wozu er Rundungsregeln benutzt. Es seien nunmehr die Relation P_1 durch die Relationsbehauptung "x und y ergeben dieselbe Zahl bei der Abrundung durch den Computer", P_2 durch die Relationsbehauptung "x und y sind gleich" definiert. Die Menge B bestehe aus allen für die Zwecke dieses Computers abgerundeten reellen Zahlen. Q sei die Relation der Gleichheit in dieser Menge B. Wiederum ergibt sich ein Strukturmorphismus von $(A; P_1, P_2)$ in $(B; Q)$ in unserem Sinne auf die folgende Weise: Jeder reellen Zahl aus A wird die zugehörige abgerundete Zahl aus B und *beiden* Relationen P_1, P_2 die Relation Q zugeordnet.

Wir können noch allgemeinere Typen von Strukturmorphismen definieren,

wenn man zwei Strukturkomplexe $(\dots, M, \dots; \dots, R, \dots)$ und $(\dots, N, \dots; \dots, S, \dots)$ mit jeweils mehreren zugrundegelegten Trägern und jeweils mehreren betrachteten Relationen auf die folgende Weise in Beziehung setzt: Als erstes ordne man jeder Menge M des ersten Komplexes, also jeder Originalmenge, als Ganzes eine der Mengen N des zweiten Komplexes, also eine Bildmenge, zu, wobei ebenfalls nicht gefordert wird, daß auf diese Weise zwei verschiedene Originalmengen unbedingt mit zwei verschiedenen Bildmengen in Verbindung gebracht werden. Zweitens führe man eine Abbildung von jeder Originalmenge in die ihr im ersten Schritt zugeteilte Bildmenge ein. Drittens ordne man jeder Relation R der Originalstruktur eine Relation Q der Bildstruktur derart zu, daß zwei Bedingungen erfüllt sind: Einmal muß wieder die Stellenzahl beider Relationen gleich sein; und zum anderen muß, wenn sich eine Variable in der Relationsbehauptung für eine Relation R auf eine bestimmte der Originalmengen M bezieht, die entsprechende Variable in der Relationsbehauptung für die diesem R zugeordnete Relation S sich auf die diesem speziellen M zugeordnete spezielle Bildmenge N beziehen. Die Begriffe Original- und Bildelemente sowie Original- und Bildrelationen werden entsprechend wie früher definiert. Folgt nunmehr für alle Relationen, die zwischen den Elementen des ersten Strukturkomplexes definiert sind, aus ihrer Gültigkeit für Originalelemente stets zwangsläufig die Gültigkeit der entsprechenden Bildrelation für die entsprechenden Bildelemente, so wollen wir auch in dieser sehr allgemeinen Situation noch von einem Morphismus sprechen.

8. DIE STRUKTUR EINES AUSSCHNITTES DER WELT

Auch wer diese letzten Gedankengänge nicht ganz nachvollziehen mag, weil in ihnen zu viele Zuordnungen gleichzeitig betrachtet werden,

wird sicherlich erahnen, daß Strukturmorphismen geeignet sind, Denkvorgänge zu beschreiben oder zu approximieren, welche vorgenommen werden, wenn man Organisationsmuster in einem Ausschnitt der realen Welt aufsucht.

Zunächst möge unter *der* Struktur eines Ausschnittes der Welt oder auch eines gedanklichen Gebildes (also z.B. der Struktur des Gehirns, der Struktur eines Unternehmens, der Struktur einer Sprache, der Struktur einer Wissenschaft usw.) die Gesamtheit *aller* Gesetzmäßigkeiten dieses Ausschnittes der Welt oder dieses abstrakten Gebildes verstanden werden, d.h. etwa eine Zusammenfassung der Form $(M_1, M_2, \dots; R_1, R_2, \dots)$, welche sich aus den Gesamtheiten der Mengen M_i , die zu diesem Ausschnitt gehören, und der Gesamtheit aller Beziehungen R_j zwischen Elementen dieser Mengen zusammensetzt. Allerdings muß man hierbei erwarten, daß - außer in ganz leicht zu übersehenden Fällen - noch erheblich allgemeinere Arten von Beziehungen als die hier bisher skizzierten Relationen benötigt werden, z.B. Begriffsbildungen, wie sie die Mathematik in topologischen oder wahrscheinlichkeitstheoretischen Strukturen definiert, doch vermutlich auch noch neu zu erfindende Begriffsbildungen, die der heutige Mathematiker noch nicht kennt und die speziellen Strukturuntersuchungen bei Forschungen in irgendwelchen Wissensgebieten entspringen können. (Man kann tatsächlich vermuten, daß durch Entdeckung neuartiger komplizierter Gesetzmäßigkeiten, z.B. in den biologischen Wissenschaften, sehr fruchtbare Impulse für die Untersuchung der dadurch angeregten mathematischen Begriffsbildungen ausgehen, d.h. z.B. neuartige mathematische Gebiete entstehen können.)

9. MATHEMATISCHE MODELLE VON EMPIRISCHEN STRUKTUREN

Daß man nun i.a. nicht *sämtliche* Gesetzmäßigkeiten eines Ausschnittes der Welt auffinden und sprachlich beschreiben kann, dürfte recht einleuchtend sein: Allein die exakte Beschreibung der zugrundeliegenden Mengen - ganz zu schweigen von den Beziehungen zwischen diesen Mengen - kann faktisch und semantisch schwierig und unmöglich sein. Ernste Streitfragen mögen entstehen, ob eine solche Beschreibung vollständig ist oder sogar, ob diese erstrebte Vollständigkeit überhaupt prinzipiell erreicht werden kann. Außerdem sind Strukturkomplexe häufig dynamischer Natur, d.h. sie verändern sich, und so ist es oft zwingend erforderlich, den Zeitpunkt oder Zeitraum genau festzulegen, auf den man die Beschreibung festlegen will. Man wird sich oft auf nicht ganz klar abgegrenzte, offene Bezugsmengen sowie auf *einige* der von dem Sachverhalt her möglichen Relationen (oder allgemeineren Arten von Beziehungen), auf eine *o f f e n e* *T e i l s t r u k t u r*, beschränken müssen. Konstruiert man nun unter diesen erschwerten Bedingungen eine Strukturabbildung eines Ausschnittes der empirischen Welt in eine mathematische Struktur, so wird man auch dann nach einem Morphismus in der hier definierten Form bezüglich sämtlicher ausgewählten und in Verbindung gebrachten Relationen (oder allgemeineren Beziehungen) streben. Falls man einen solchen Morphismus inständig erhofft oder dieser mehr oder weniger genau vorliegt, nennt man die Strukturabbildung ein *m a t h e - m a t i s c h e s M o d e l l* der empirischen Struktur ([4], S. 125). (*Häufig wird auch nur die mathematische Bildstruktur als Modell bezeichnet*, wobei dann die Originalmengen und Originalrelationen und die verbindenden Zuordnungen *stillschweigend als bekannt* vorausgesetzt werden.) Derartige Modelle können zu deskriptiven oder präskriptiven Zwecken aufgestellt werden, d.h. um die empirische

Struktur zu beschreiben und analysieren zu können oder um sie beeinflussen und regulieren zu können.

Es kann nun nicht genug betont werden, daß die moderne Mathematik weit mehr Hilfsmittel als das Rechnen mit Zahlen und Zahlenfolgen und das Hantieren mit geometrischen Figuren zur Hand hat, um Probleme in solchen mathematischen Bildstrukturen zu lösen, die Fragestellungen in der empirischen Struktur entsprechen. Z.B. lassen sich die eng miteinander verknüpften Methoden der Relationentheorie, mathematischen Logik, Graphentheorie und Booleschen Matrizentheorie in immer vielfältigerer Weise auf Probleme anwenden, welche Fragestellungen aus solchen empirischen Gebieten widerspiegeln, bei denen quantitative Begriffe noch relativ wenig entwickelt und untersucht sind (z.B. bei einigen Gesellschaftswissenschaften, siehe [4]).

Die Anwendungsmöglichkeiten von mathematischen Modellen sind äußerst vielseitig und breiten sich auf immer mehr Gebiete aus. Neben recht allgemein bekannten Modellen aus den Natur- und Ingenieurwissenschaften haben sich ja inzwischen z.B. auch Modelle als außerordentlich wertvoll erwiesen, mit denen man Vorhaben plant, Ketten logischer Schlußfolgerungen nachvollzieht und ihre logische Stichhaltigkeit prüft, komplexe Folgerungen aus einfachen Grundsätzen ableitet, Sprachen analysiert, gegen die Luftverschmutzung und gegen Krebs kämpft, gegen oder für Zigarettenkonsum wirbt, Weltraumflüge simuliert usw.

10. DER PROZESS DER MODELLBILDUNG

Der Prozeß der Modellbildung geht nun, etwas ausführlicher beschrieben, folgendermaßen vor sich: Ein Ausschnitt der realen Welt werde zunächst verbal beschrieben. Bereits dieser erste Prozeß ist eigent-

lich ein Abbildungsvorgang der oben beschriebenen Art. Ausgangspunkt ist i.a. ein außerordentlich verwickelter Strukturkomplex, den wir empirisch erkennen. Der Versuch der Beschreibung ergibt zunächst eine verbal formulierte Struktur, die auf unseren Abstraktionen basiert, wobei wir erstreben, daß man durch diese Abbildung einen Strukturmorphismus gewinnt. Daß jedoch bereits dieser Übergang i.a. keinen Strukturisomorphismus ergibt, ist in Kreisen der Semantiker wohlbekannt. Diese sprechen von dem **G r u n d s a t z d e r N i c h t - A l l g e m e i n h e i t**, welcher den wesentlichen Unterschied zwischen Wirklichkeit und Beschreibung der Wirklichkeit betont. (Gleichgültig wie viel man über eine "Sache", einen "Vorgang", eine "Eigenschaft oder irgendetwas anderes" aussagt, man kann i.a. nicht alles darüber aussagen. Selbst ganz einfache Aussagen sind nur innerhalb gewisser Grenzen wahr. Siehe [3], S.16.)

Ausgehend von dieser ersten, semantisch bestimmten Struktur suchen wir dann durch Abstraktion oder durch eine Folge von Abstraktionen, die wir gedanklich zu einer Gesamt-Abstraktion zusammenschalten können, zu einer weiteren Struktur derart überzugehen, daß möglichst ein Strukturmorphismus bei dem Übergang gegeben ist. Dabei erhofft man immer noch, daß über die Kompatibilität hinaus ein Strukturisomorphismus besteht, d.h. daß man von den Bildern zu den Originalen und von den Relationsbehauptungen in den Bildstrukturen zu den Relationsbehauptungen in den Originalstrukturen auf eindeutige Weise zurückkehren kann.

Aber leider ist eine solche Hoffnung auch in ganz einfachen Fällen recht fragwürdig. Im Grunde genommen bewegen sich nämlich unsere Aussagen nur in einem abgeschlossenen System der Abstraktionen, die wir geprägt haben, und nicht in dem offenen System der Wirklichkeit.

Dieser Erkenntnis sind sich die Menschen oft nicht bewußt, so daß die Abstraktionen mit der Wirklichkeit verwechselt werden. Wenn plötzlich die Erleuchtung auftritt, daß Aussagen sich nicht auf die Wirklichkeit, sondern nur auf Abstraktionen beziehen, so kann leicht als eine Folgerung aus dieser plötzlichen Erkenntnis der Vorwurf der "Manipulation" entspringen. Ein typisches Beispiel hierfür ist die oft gehörte Klage: "Statistiken lügen!" (obwohl natürlich auch hier, ebenso wie bei allen anderen Typen von Aussagen, die Möglichkeit einer Manipulation vorliegen kann, nämlich durch bewußt einseitige, tendenziöse Auswahl der Informationen).

11. DIE UNVOLLSTÄNDIGKEIT VON MATHEMATISCHEN MODELEN DER WIRKLICHKEIT

Zum Schluß seien noch einmal die prinzipiellen Schwierigkeiten einer Idealisierung eines Ausschnittes der Wirklichkeit, wie sie bei einer Modellbildung vorgenommen wird, kurz zusammengestellt.

1. Die empirische Originalstruktur ist möglicherweise nicht klar abgrenzbar oder nicht intersubjektiv verständlich beschreibbar.
2. Die einzelnen Abbildungen sind möglicherweise nur scheinbar, nur bei oberflächlicher Betrachtung oder nur beim heutigen Stand unserer Erkenntnis Morphismen.
3. Ob bei einer solchen Abbildung ein Isomorphismus vorliegt, ist oft sehr fraglich.

Bei einer Rückübersetzung der Erkenntnisse, welche über die mathematische Bildstruktur gewonnen wurden, in Informationen über die Wirklichkeit können daher drei wesentliche Fehlerquellen entstehen:

1. Die Wirklichkeit umfaßt mehr Originalelemente oder andere Originalrelationen, als aus den mathematischen Bildern ersichtlich ist.

2. Die Ähnlichkeit zwischen Wirklichkeit und mathematischem Bild ist unzureichend.
3. Die abstrakte Gleichheit zwischen Wirklichkeit und mathematischem Bild ist eine Fiktion.

Daß man trotz dieser gravierenden Einschränkungen vom Modell recht fruchtbare Rückschlüsse auf die Wirklichkeit ziehen kann, zeigen die Erfolge der Praxis. Der erfolgreiche Praktiker kennt nämlich die Grenzen seiner Modelle und hütet sich, diese zu überschreiten.

LITERATUR

- [1] H. BEHNKE, R. REMMERT, H.-G. STEINER, H. TIETZ (Herausgeber):
Das Fischer Lexikon der Mathematik 1.
Fischer-Bücherei, Frankfurt 1966
- [2] G.B. DANTZIG:
Lineare Programmierung und Erweiterungen (Deutsche Bearbeitung von A. Jaeger).
Springer-Verlag, Berlin-Heidelberg-New York 1966
- [3] S.I. HAYAKAWA (Herausgeber):
Wort und Wirklichkeit.
Verlag Darmstädter Blätter, Darmstadt 1970
- [4] A. JAEGER, K. WENKE:
Lineare Wirtschaftsalgebra.
Teubner, Stuttgart 1969
- [5] W. LEINFELLNER:
Einführung in die Erkenntnis- und Wissenschaftstheorie.
Bibliographisches Institut, Mannheim 1965

**Numerical Aspects of Pollution Abatement Problems:
Optimal Control Strategies for Air Quality Standards**
by K. O. Kortanek and W. L. Gorr, Pittsburgh

ABSTRACT

The Environmental Protection Agency, U.S.A., has proposed National Air Quality Standards for classes of air pollution, designed to protect public health and welfare by setting limits on levels of pollution in the air. They apply to all areas of the United States. Air quality standards give rise to an infinite number of restrictions, called the "air quality constraint set," on the ambient air quality of a region; that is to say, the concentration at every ground level point must be less than or equal to the standards on the average.

Diffusion modeling bridges the gap between emissions and ambient concentrations and permits specification of the air quality constraint set. We use one particular form of diffusion model, by drawing on its algebraic properties, which it also shares with many diffusion models.

We present an approach to determine regional emission regulations that are (1) feasible in the sense that they lead to compliance with air quality standards and (2) optimal in the sense that the total economic impact due to implementation is minimized. Under goal (1), we set forth a procedure more sophisticated than heuristic procedures such as those guaranteeing compliance on a fixed rectangular set of grid receptor points in a region.

Our procedure does not presuppose a finite set of constraints arbitrarily set in advance, but the correct finite set is determined by the entire model and the data. The model will determine the location of a set of maximum receptor points in the region, which then will be the optimal

location of sampling stations. If compliance is achieved at these strategic points in the region during the designated time interval, then compliance is achieved at all points in the region with respect to an optimal air quality standard.

1. Introduction

The approach to pollution abatement by imposition of air quality standards appears to be one way in which the Environmental Protection Agency (EPA) is following. A standard specifies a level of concentrations of pollutant which is not to be exceeded on the average within a prescribed period of time. It appears that a prescription of standards must hold at all ground level (two-dimensional) coordinates. Current diffusion models appear to fall into two main categories--short-term or episodal, and long-term. Most of the diffusion models in the literature are based on a solution (statistical in nature) to an initial-value, boundary-value problem which yields a spatial distribution of a pollutant within the class of named multivariate probability distributions. They are differentiated through assumptions as to the type of emission source, functional complexity of the parameters, and the nature of the iteration method employed. See Herzog [33] and also Sutton [51], Pasquill [43], [44], Gifford [25], Turner [54], Martin-Tikvart [40], Fortak [23], Roberts, et al [46], and others.

Normally, pollution policies are based on a few dominating sampling points in a region, which may first be divided into a grid of areas. The assumption then is that if pollution is below a standard at the sampling stations, then it is below the standard at all points in the region. Thus it appears that the region requires that emission rates of pollutant sources be below certain levels, demanding a fortiori that the sources take whatever measures necessary that appear economically feasible.

It is the case that the 1967 Clean Air Act and the new 1970 Clean Air Amendments [12] are stated in terms of air quality standards. Therefore, part of pollution abatement efforts must rely heavily on diffusion modeling.

One of the most important uses of input data for a mathematical multiple-source urban air pollution model is a real time short-term prediction of concentration for the entire urban area (see Fortak [23] and others).

However, as Professor Fortak emphasizes, the problems of city planning are connected with the problem of simulation and prediction of long-term climatological ground-level concentration fields. "Here, time series of concentrations are of minor interest; instead statistics of observed or calculated concentration fields for given long periods of time are important (Fortak [23], page 9-2).

In this paper we develop an optimization model involving estimations for the annual ground level pollution due to area or point sources of pollution involving approximately unimodel pollution surfaces of the kind referenced above. Using as decision variables the weight fraction of SO_2 and particulate removed at the sources, we illustrate this class of models by a single objective function of minimizing total cost from all sources while keeping total emissions below standards at all coordinates of the region. Since both short and long-term have been related implicitly by Legislation (see [12]), we also develop a selective abatement procedure model which, for the first time, considers short and long-term standards simultaneously.

We therefore achieve a model formulation which is consistent with any functional relationship whatever, and in particular, well suited when such relationships are based on a single parameter such as time, t . No longer is the model dependent on the grid size, or the space between grids. In fact, we prove that only a finite number of points of concentration in the region need enter an optimal solution, as ground level coordinates for

the basic pollution surfaces. Thus, by appealing to recent results on generalized moment problems [29], [30], we render a continuous parameter problem into an equivalent one with only a finite number of parameter values needed.

As already stated by several authors, the functional complexity of the parameters and the iteration method employed are intrinsic parts of diffusion modeling. It appears, however, that the literature has not yet recognized that optimization problems stemming from models such as the Martin-Tikvart, Fortak or Roberts, et al. models are instances of semi-infinite programming problems [6], [7], [8]. Once this identification is made, it is possible to give a generalized moment type problem whose solution is equivalent to solving a (semi-infinite) Martin-Tikvart problem. In fact, following Gustafson [29] and Gustafson-Kortanek-Rom [31], it is seen that a generalized moment problem is "dual" to the Martin-Tikvart problem. We thus develop models which display the functional complexity of parameters, and incorporate efficient iteration methods of solution, namely algorithms for solving generalized moment problems as developed in [29] and [31], based on Newton-Raphson subroutines and implementable via computer codes.

2. Pollution Abatement Problem (PAP) Formulations

2.1 A Two-Dimensional Minimum Cost Diffusion Model

We cite an optimization problem frequently found in the literature, as presented for example by Mr. Wilpen Gorr [26], stemming from the work of Martin-Tikvart [40].

The annual average ground level pollution due to N stationary sources is calculated as:

$$(2.1) \quad X_T(x,y) = \sum_{j=1}^N Q_j \psi_j(x,y) \quad \text{where}$$

$$(2.2) \quad X_T = \text{annual average concentration at ground level coordinate } (x,y) \text{ in units of } \mu\text{gm/m}^3,$$

$$(2.3) \quad Q_j = \text{present average emission rate of pollutant in mgm/sec at source } j, \text{ and}$$

$$(2.4) \quad \psi_j = \text{basic pollution surface for source } j, \text{ approximately unimodal.}$$

Decision variables are defined by:

$$(2.5) \quad E_{S_j} = \text{weight fraction of } \text{SO}_2 \text{ removed at source } j$$

$$(2.6) \quad E_{P_j} = \text{weight fraction of particulate removed at source } j.$$

The semi-infinite optimization problem (finite number of variables, finite number of constants) becomes in this notation:

$$\min \sum_{j=1}^N c_{S_j} (E_{S_j})$$

$$(2.7) \quad \text{subject to} \quad \sum_{j=1}^N (1 - E_{S_j}) Q_{S_j} \psi_j(x, y) \leq S, \text{ for all } (x, y) \in A$$

$$E_{S_j} \geq E_{S_j}^L$$

$$- E_{S_j} \geq - E_{S_j}^U \quad 1 \leq j \leq N.$$

In this problem we use:

$$(2.8) \quad Q_{S_j} = \text{SO}_2 \text{ emission rate at source } j$$

$$(2.9) \quad c_{S_j}(E_{S_j}) = \text{cost of attaining } E_{S_j}$$

$$(2.10) \quad S = \text{annual SO}_2 \text{ air quality standard}$$

$$(2.11) \quad A = \text{air pollution control region (county, etc.)},$$

$$(2.12) \quad E_{S_j}^L, E_{S_j}^U = \text{lower and upper bounds on the } E_{S_j} \text{ - variables.}$$

Observe that (2.7) may be further refined by functional relationships involving a single parameter such as time, t . For example, x and y coordinates may be converted to functions of time $x(t)$, $y(t)$ where distances are measured along trajectories.

Note also that a similar optimization problem may be given for particulates involving decision variables E_{P_j} , either by itself or mixed with problem (2.7).

Observe that problem (2.7) has as many linear constraints as there are coordinate points in the control region, namely an infinite number of them. However, the number of variables is finite, and therefore problem (2.7) is a semi-infinite programming problem of the type developed and studied by Charnes-Cooper Kortanek [7], [8], and [9].

2.2 A One-Dimensional Analog of Problem (2.7)

In order to relate recent results on generalized moment theory of Gustafson, Kortanek and Rom to PAP, we consider a one-dimensional analog of optimization problem (2.7), where for convenience we use a linear objective function.¹⁾

$$\begin{aligned}
 (2.2.1) \quad & \min \sum_{j=1}^N c_{S_j} E_{S_j} \\
 & \text{s.t.} \quad \sum_{j=1}^N \left(1 - E_{S_j}\right) Q_{S_j} \psi_j(t) \leq S, \quad \text{for all } t \in I \\
 & E_{S_j} \geq E_{S_j}^L \\
 & -E_{S_j} \geq -E_{S_j}^U, \quad 1 \leq j \leq N.
 \end{aligned}$$

Here I is a closed interval, taken for example to be $[0,1]$. The N sources are at distinct points in I . The functions $\psi_j(t) = e^{-(t-\delta_j)^2}$ where source j would be at the point δ_j in I , are examples of unimodal surfaces of normal distribution type.

2.2.2 A Restatement of Problem (2.2.1) in Convenient Notation

Continuing our goal of relating problem (2.2.1) to moment theory, we write problem (2.2.1) in the following, equivalent form:

$$\inf \sum_{r=1}^N y_r \mu_r$$

¹⁾ Formulation (2.2.1) may be equivalent to a two-dimensional one via certain general functional relationships. In some cases, therefore, (2.2.1) may not in effect be a one-dimensional problem initially.

$$(2.2.2.1) \quad \text{subject to } \sum_{r=1}^N y_r u_r(t) \geq \phi(t) \quad \text{for all } t \in [0,1]$$

$$y_r \geq y_r^L$$

$$-y_r \leq -y_r^U \quad 1 \leq r \leq N$$

Now the y_r 's are the variables, μ_r 's the cost, and $u_r(t)$ are continuous functions on I . Here "inf" replaces "min" because in the general theory, an infimum may not be a minimum. Problem (2.2.2.1) is analogous to problem D in [29] and [31], but with bounded variables. These extra bounding conditions give a more general problem than D, but one which can be handled by extensions of their numerical methods.

2.2.3 The Moment Mathematical Model - See [29]

For numerical solution of problem type (2.2.2.1), we actually propose to numerically solve its "dual" problem. To this end we turn now to [29] and [31] for development of the dual moment problem.

Let E be the linear space of functions α which are of bounded variation on $[a,b]$, continuous to the right in (a,b) , and normalized by the condition $\alpha(a) = 0$. We introduce the norm $\|\alpha\|$ defined by

$$\|\alpha\| = \int_a^b |d\alpha(t)|$$

Thus E is a Banach space.

Let F be the linear space of functions, continuous on $[a,b]$ denoted by $C[a,b]$ and normed by

$$\|f\| = \max_{t \in [a,b]} |f(t)|$$

Thus F is a Banach space.

If L is a bounded linear functional on $C[a,b]$ then there exists an α in E which is independent of f and such that

$$L(f) = \int_a^b f(t) d\alpha(t)$$

We therefore can introduce the pairing \langle , \rangle between E and F

$$\langle f, \alpha \rangle = \int_a^b f(t) d\alpha(t) \quad f \in F, \alpha \in E.$$

Denote by R^n the Euclidean n -dimensional vector space. Choose n functions u_1, u_2, \dots, u_n from F and define the operator $U: E \rightarrow R^n$ by $U(\alpha) = y$ where

$$\int_a^b u_r(t) d\alpha(t) = y_r, \quad r = 1, 2, \dots, n$$

Now let $\bar{\Phi}$ be a fixed function in F and μ a fixed vector in R^n . Then we can formulate the problem

$$\begin{array}{ll} (2.2.3.1) & \text{Compute } \sup \quad \langle \bar{\Phi}, \alpha \rangle \\ & \text{subject to} \quad U(\alpha) = \mu \\ & \text{and} \quad \alpha \nearrow \end{array}$$

Let C denote the set of functions α which are not decreasing and of bounded variation over $[a, b]$. It is easily verified that C is a closed cone. We now use the procedure outlined in [29] and formulate the pair of dual problems P and D

$$\begin{array}{ll} P: \max < \Phi, \alpha > & D: \min (\mu, y) \\ \text{subject to } U(\alpha) = \mu & \text{subject to } U^T(y) - \Phi \in C^* \\ \alpha \in C & \end{array}$$

Here (\cdot, \cdot) denotes scalar product in R^n . The adjoint operator U^T of U , $U^T: R^n \rightarrow F$ is defined by the requirement

$$(2.2.3.2) \quad (U(\alpha), y) = < \alpha, U^T y >$$

for all α in C and y in R^n . It is easily proved that

$$U^T y = \sum_{r=1}^n y_r u_r$$

and that this is the only way of satisfying (2.2.3.2). Next we define C^* by the relation

$$C^* = \{f: f \in F, \alpha \in C, < f, \alpha > \geq 0\}$$

We immediately find

$$C^* = \{f: f \in F, f(t) \geq 0, \forall t \in [a, b]\}$$

Now we can formulate P and D in the following way

$$\begin{aligned}
 (2.2.3.3) \quad P: \quad & \sup \int_a^b \phi(t) d\alpha(t) & D: \quad & \inf \sum_{r=1}^n y_r \mu_r \\
 \text{subj. to} \quad & \int_a^b u_r(t) d\alpha(t) = \mu_r, & \text{subj. to} \quad & \sum_{r=1}^n y_r u_r(t) \geq \phi(t), t \in [a, b] \\
 & \alpha \nearrow & & \\
 & r = 1, 2, \dots, n & &
 \end{aligned}$$

Problem (2.2.2.1) is a slight extension of problem D in Gustafson [29], having the additional bounding conditions on the y_r -variables.

2.3.4 Some Properties Leading to an Algorithm. See [29]

Gustafson [29] and Gustafson-Kortanek-Rom [31] consider regularized versions of problem D (denoted by D_F) where additional redundant constraints $|y_r| \leq F$ are adjoined. These are introduced for purposes of solution rather than being part of the problem structure.

Hence the moment problem associated with (2.2.2.1) becomes:

$$\begin{aligned}
 (2.3.4.1) \quad & \sup \int_0^1 \phi(t) d\alpha(t) + \sum_{r=1}^n (y_r^L v_r^L - y_r^U v_r^U) \\
 \text{s.t.} \quad & \int_0^1 u_r(t) d\alpha(t) + v_r^L - v_r^U = \mu_r, r = 1, 2, \dots, n \\
 & \alpha \nearrow v_r^L, v_r^U \geq 0
 \end{aligned}$$

The following Theorems are related to analogous ones in [29] and [31]. Their proofs follow analogously also, providing one replaces $|y_r| \leq F$ by the pair $u_r \geq y_r^L$, $y_r \leq y_r^U$ for each variable r .

Theorem 1 (See [29]) Assume some $u_j(t) > 0$ ¹⁾ for all $t \in [0,1]$, and that $\max_{t \in [0,1]} |\dot{\phi}(t)|/u_j(t) < y_j^U$. Then only two cases for problems (2.2.2.1) and (2.3.4.1) are possible:

1. Both are consistent and have bounded solutions
2. (2.3.4.1) is inconsistent and (2.2.2.1) has unbounded solutions.

Theorem 2 The objective value of (2.3.4.1) can be taken on for an integrator with at most n points of increase, i.e. a step function.

Remark It is this theorem which permits replacement of continuously parameterized problem by a discrete one. However, there is no apriori reason for the mass points to be "evenly spaced" such as the case for a grid system.

The following theorem yields a system of nonlinear equations for the n points of increase which leads to an algorithm (see [29], [31]) presented in the appendix.

¹⁾ This will be true for $u_1(t)$ if the system $\{u_i(t)\}_{1 \leq i \leq n}$ is a Chebyšev system, see [28].

Theorem 3 (See [29]) Assume an optimal solution of (2.3.4.1) is given by:

- (i) mass ρ_i at point $t_i \in [0,1]$, for $i = 1, 2, \dots, q$
- (ii) q^U of the $v_r^U > 0$, namely $v_{r_1}^U, \dots, v_{r_q}^U$
- (iii) q^L of the $v_r^L > 0$, namely $v_{s_1}^L, \dots, v_{s_q}^L$

Let y_1, \dots, y_n be an optimal solution of (2.2.2.1). Then the following equations are satisfied:

$$(T3.1) \quad \sum_{i=1}^q \rho_i u_r(t_i) + v_r^L - v_r^U = \mu_r ; \quad r = 1, 2, \dots, n$$

$$(T3.2) \quad \sum_{r=1}^n y_r u_r(t_i) = \phi(t_i) ; \quad i = 1, 2, \dots, q$$

$$(T3.3) \quad y_{r_j} = y_j^U \quad j = 1, 2, \dots, q^U$$

$$(T3.4) \quad y_{s_j} = y_j^L \quad j = 1, 2, \dots, q^L$$

$$(T3.5) \quad \sum_{r=1}^n y_r u'_r(t_i) = \phi'(t_i) \quad i = 1, 2, \dots, q$$

These necessary conditions lead to an algorithm essentially identical to [29] and [31].

As is apparent from [31], several different states can occur, which are not encountered in finite linear programming. Actually, a variety of analytical problems arise depending on assumptions imposed on the functions u_1, u_2, \dots, u_n . For example, if they form a complete Chebyshev system, then the results are simpler but still permit the treatment of a

wide range of problems. An algorithm for numerical solution of the generalized moment problem for one dimension has been developed in Gustafson-Kortanek-Rom [31], Gustafson-Rom [32] and Gustafson [29].

3 Conditional-Adaptive Planning

In industrial engineering contexts such as multi-product production scheduling [36], [49], the sequential nature of production planning has been explicitly taken into account. Long- and short-term horizons have been designated, with the short-term decisions being implemented and the long-term providing overall guidance. By implementing the short-term plan, we increase the chance of a quicker response to change while implicitly incorporating the intertemporal character of the entire horizon (both short and long) to shape the current period decision. In fact, by using constraint oriented devices such as "horizon-postures" recursively, one may indirectly minimize changeover and start-up effects across all horizons.

In pollution abatement, it appears that both short and long-term horizons have been related via a standards approach. Short-term standards impose higher concentrations vis-a-vis long term ones. On the other hand, however, very high short-term stagnation pollution potentials have resulted in temporary and economically expensive control measures such as fuel switching, cutting back on operations, etc.

The problem of designing emission limitations for both short and long-term air quality standards can be a difficult one. Short-term diffusion modeling appears to be a much less precise tool than long-term diffusion modeling, the reason being statistical (law of averages) in nature. In some cases, such as with SO_2 , it appears that the long term air quality standards are dominating in the sense that PAP-designed regulations would also guarantee compliance with 24-hour standards. Here one can demonstrate with simple calculations that the effect of "episode" curtailments of emissions has negligible effect on annual average concentrations. Thus the PAP formulation is applicable.

On the other hand, fine particulate standards appear to have the short term requiring the more restrictive regulations. In this case one might redefine the annual standard using one of the many empirical short term/long term concentration relationships so that the redefined standard dominates the short term standard. Again the PAP formulation would be relevant in a practical way since we can more readily model annual concentrations.

It appears at this point that the issues of changeover, horizons and conditional-adaptive planning are ones which cannot be avoided if there is interest in numerical results including implementation. We are interested in developing "laws" or "functional relationships" wherever possible in order to simplify complex relationships. We cannot wait, however.

for the complete master plan which delineates the function task of each entity or "grid cell" of the city. We intend to develop analytical tools which assist in prediction and control of a system without complete dependency on perfect understanding, as detailed by Professors Cooper, Eastman, Johnson and Kortanek, e.g. see the section "A. The Modeling Effort," pp. B742-B745 of [48].

4. Optimum Emission Control Strategies Achieving Air Quality Standards

The current best known methods for the formulation and evaluation of emission control strategies (see [22], [52] and [53] below) involve searching through a prespecified list of strategies to see which of these meet air quality standards at all grid points in a region, and which minimize total regional cost. These methods are trial and error procedures and involve the computer through combinatorial simulations. Such heuristic procedures (see [53]) appear to be of real value in getting a start on the problem of establishing implementation plans. The essential ingredients of the method are: (1) have the user specify various emission standards for particular pollutants in a region, (2) for each of these compute the changes in pollution surfaces and test for compliance with air quality standards at a finite set of receptor points in the region, and (3) of those strategies which do comply at these points, select one which minimizes total regional cost and is technically feasible with respect to available devices.

We view our development on optimal control strategies to be undertaken along with these. We have a more sophisticated procedure to be sure, but one which could be used to measure the effectiveness of the heuristic procedure, including justifying those instances where the heuristics alone are enough. The optimal procedure could single out instances, however, where additional refinements are necessary, thereby reducing the chances of embarrassing consequences from a straightforward application of the heuristic approach to all cases.

5. Some Methods for Further Restrictions on the E_j -Reductions: Towards Regulatory Policies [39]¹⁾

5.1 Min-Max Percent Reduction per Source

Since we have already indicated elsewhere preliminary responses to potential problems that least cost strategies may impose for regulatory policy, [39], we indicate some ways in which our technology could embrace modifications in order to obtain a more accurate picture of the total impact of implementation of control strategies. The E_j -variables introduced in section 2 are the weight fraction of SO_2 removed at source j , E_{S_j} , or the weight fraction of particulate removed at source j , E_{P_j} . These variables yield selective abatement procedures, and lead to regional emission regulations incorporating possibly different reductions for different sources.

¹⁾ We would like to acknowledge helpful discussions with Professors Blumstein, Cassidy and Walters of the School of Urban and Public Affairs. See also reference [1].

The numerical values of these emission reductions are determined that are (1) feasible in the sense that they lead to compliance with air quality standards throughout the region and (2) optimal in the sense that the "total" economic impact due to implementation is minimized.

In section 2 we introduced lower limits and upper limits for each source reduction E_j , which would be tuned in practice according to a host of attributes of particular sources depending on the region in which they are located, local politics, etc. Thus, we require that each E_j be bigger than or equal to a certain lower limit E_j^L , and smaller than or equal to an upper limit, E_j^U . In terms of a scenario at this time, an agency could specify the min-max reduction percentages within which we would seek emission reductions, E_j , which would be now the least cost, but from among those meeting goal (1).

It is probably the case that this modification is still too fine for practice. However, it does lead into the following modification based on classifications of source types.

5.2. Competitive n-Person Game Relations Between the E_j 's

In order to not rely exclusively on a "price" system to distribute total regional impact among sources, we can introduce various competitive game theoretic concepts which involve dominance and stability (including political). Included in this approach would be a weighting procedure among the E_j -classes which might involve a regulator's estimate

of a mutually stable relationship among sources in a region. Another concept could involve the notion of "undominated" emission regulations to be determined so that one source cannot develop a "justified objection" against any of the other sources.

In any of the modifications (1) through (4), our available technology will yield systems measures for the cost-benefit consequences of their imposition. For example, we can obtain the change in "total regional cost" as measured one particular way before the introduction of a regulatory modification and afterwards. In this way, we can get a measure on possible inequities and increased control costs above and beyond the initial way chosen to measure "total regional cost," as for example, set forth in our models in sections 2 and 3.

In summary, any of these modifications suggested by legal regulatory considerations would be goals required to be met in addition to (1), while attempting to seek (2), an optimal least cost solution under the altered circumstances. It is because of additional constraints and restrictions that we have dubbed our numerical solution methods, "constrained generalized moment techniques."

BIBLIOGRAPHY

- [1] Blumstein, A., R. G. Cassidy, W. L. Gorr and A. S. Walters, "Optimal Specification of Emission Regulations Including Reliability Requirements," Paper presented to the 1971 Annual Conference of the Canadian Operations Research Society, Ottawa, June 1971.
- [2] Blumstein, A., R. W. Dunlap and A. Walters, "Course Syllabus: Physical Technical Processes," School of Urban and Public Affairs, Carnegie-Mellon University.
- [3] Branch, M. C., "Delusions and Diffusions of City Planning in the United States," Management Science 16, August 1970, B714-B732.
- [4] Byrne, R., A. Charnes, W. W. Cooper and K. O. Kortanek, "A Chance-Constrained Programming Approach to Capital Budgeting with Portfolio Type Payback and Liquidity Constraints and Horizon Posture Controls," Journal of Financial and Quantitative Analysis, 2, No. 4, December 1967, pp. 339-364.
- [5] Charnes, A. and W. W. Cooper, Management Models and Industrial Applications of Linear Programming (J. Wiley and Sons, New York: 1961) Vols. I and II.
- [6] Charnes, A., W. W. Cooper and K. O. Kortanek, "Duality, Haar Programs and Finite Sequence Spaces," Proc. Nat. Acad. Sci. U.S., 48 (1962) pp. 783-786.
- [7] Charnes, A., W. W. Cooper and K. O. Kortanek, "A Duality Theory for Convex Programs with Convex Constraints," Bull. Amer. Math. Soc., 68 (1962) pp. 605-608.
- [8] Charnes, A., W. W. Cooper and K. O. Kortanek, "Duality in Semi-Infinite Programs and Some Works of Haar and Caratheodory," Management Science, 9 (1963) pp. 209-228.
- [9] Charnes, A., W. W. Cooper and K. O. Kortanek, "On the Theory of Semi-Infinite Programming and a Generalization of the Kuhn-Tucker Saddle Point Theorem for Arbitrary Convex Functions," NRLQ, 16 (1969) pp. 41-51.
- [10] Cheney, W. E., Introduction to Approximation Theory (McGraw-Hill, Inc.: 1966).
- [11] Clarke, J. F., "A Simple Diffusion Model for Calculating Point Concentrations from Multiple Sources," J. Air Pollution Control Assoc., 14 (September 1964) pp. 347-352.
- [12] "Clean Air Amendments of 1970," Public Law 91-604, 91st Congress, H.R. 17255, December 31, 1970.
- [13] Conway, R. W. and W. L. Maxwell, Theory of Scheduling (Addison Wesley: 1967).

- [14] Cooper, W. W., C. M. Eastman and K. O. Kortanek, "Systems Approaches to Urban Planning: Mixed Conditional Adaptive and Other Alternatives," Institute of Physical Planning, No. 6, C-M University, August 1970.
- [15] Dantzig, G. B., Linear Programming and Extensions (Princeton University Press: 1963).
- [16] Davidson, B., "A Summary of the New York Urban Air Pollution Dynamics Research Program," J. Air Pollution Control Assoc., 17 (March 1967) pp. 154-158.
- [17] Day, Mahlon M., Normed Linear Spaces (Springer-Verlag, New York: 1962).
- [18] Duffin, R. J., "Infinite Programs," pp. 157-170 in: Linear Inequalities and Related Systems (ed. by H. W. Kuhn and A. W. Tucker), Annals of Math. Studies No. 38, Princeton University Press, Princeton, N.J., 1965.
- [19] Eastman, C. M. and K. O. Kortanek, "Modeling School Facility Requirements in New Communities," Management Science, 16 (August 1970) pp. B784-B799.
- [20] Eastman, C. M. and K. O. Kortanek, "Adaptive Conditional Approaches to Urban Planning," Institute of Physical Planning, No. 4, C-M University, March 1970.
- [21] Eastman, C. M., N. J. Johnson and K. O. Kortanek, "A New Approach to an Urban Information Process," Management Science, 16 (August 1970) pp. B733-B748.
- [22] Farmer, J. R., P. J. Bierbaum and J. A. Tikvart, "Proceeding from Air Quality Standards to Emission Standards," APCA Paper 70-85, presented at the Annual Meeting of the Air Pollution Control Association, St. Louis, Missouri, June 14-19, 1970.
- [23] Fortak, H. G., "Numerical Simulation of the Temporal and Spatial Distributions of Urban Air Pollution Concentration," in Proceedings of Symposium on Multiple-Source Urban Diffusion Models, Arthur C. Stern, ed., U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 1970.
- [24] Frenkiel, F. N., "Turbulent Diffusion: Mean Concentration Distribution in a Flow Field of Homogeneous Turbulence," in Advances in Applied Mechanics, eds. R. and T. von Karman, Vol. III, Academic Press, Inc., New York, 1953, pp. 61-107.
- [25] Gifford, F. A., "Diffusion in the Diabetic Surface Layer," J. Geophys. Research, 67(8) (July 1962) pp. 3207-3212.
- [26] Gorr, W., Working Papers, School of Urban and Public Affairs, Carnegie-Mellon University
- [26A] "The Single Source Compliance with Air Quality Standards Problem Including Reliability Requirements of Air Pollution Control Devices."

- [26B] "Minimization of Direct Damage and Control Costs of Air Pollution."
- [26C] "The Basic Compliance with 'Air Quality Standards' Air Pollution Abatement Problem--An Overview," September 1970.
- [27] Gorr, W. and K. O. Kortanek, "Cost Benefit Measures for Regional Air Pollution Abatement Models," Institute of Physical Planning, No. 16, C-M University, May 1971.
- [28] Gorr, W. and K. O. Kortanek, "Optimal Control Strategies and Regulatory Policy," invited lecture for the Tenth Annual German Operations Research Society, Bochum, Germany, September 1971.
- [29] Gustafson, S.-A., "On the Computational Solution of a Class of Generalized Moment Problems," SIAM J. Numer. Analysis, 7, No. 3 (1970) pp. 343-357.
- [30] Gustafson, S.-A., "Numerical Aspects of the Moment Problem," Filosofie Licentiat, Institutionen for Informations Behandling, Stockholms Universitet, Stockholm, Sweden, April 1970.
- [31] Gustafson, S.-A., K. O. Kortanek and W. Rom, "Non-Cebyseviaan Moment Problems," SIAM J. Numer. Analysis, 7, No. 3 (1970) pp. 335-342.
- [32] Gustafson, S.-A. and W. Rom, "Applications of Semi-Infinite Programming to the Computational Solution of Approximation Problems," Tech. Rep. No. 88, Department of Operations Research, Cornell University, Ithaca, N.Y., September 1969.
- [33] Herzog, H. W., Jr., "The Air Diffusion Model as an Urban Planning Tool," Socio-Economic Planning Science, 3 (1969) pp. 329-349.
- [34] Karlin, S. J. and W. J. Studden, Tchebysheff Systems: With Applications in Analysis and Statistics (Interscience Publishers, New York-London-Sydney: 1966).
- [35] Khamis, S. H., "On the Reduced Moment Problem," Ann. Math. Stat., 25 (1956) pp. 133-122.
- [36] Kortanek, K. O., D. Sodaro and A. L. Soyster, "Multi-Product Production Scheduling via Extreme Point Properties of Linear Programming, NRLQ, 15 (June 1968) pp. 287-300.
- [37] Kortanek, K. O., "Optimum Emission Control Strategies Achieving Air Quality Standards," private communication to J. A. Tikvart, January 18, 1971.
- [38] Kortanek, K. O., "Optimal Control Strategies for Air Quality Standards," NAPCA Proposal, January 26, 1971.
- [39] Kortanek, K. O., "Selective Abatement Procedures Modifications and Regulatory Policy," private communication to G. R. Perkinson (EPA), March 8, 1971.

- [40] Martin, D. and J. Tikvart, Air Quality Display Model, Department of Health, Education and Welfare, No. PH 22-68-60, Washington, D. C., November 1969.
- [41] Martos, B., "Quasi-Convexity and Quasi-Monotonicity in Nonlinear Programming," Institute of Economics of the Hungarian Academy of Sciences, No. 20, Budapest, 1966.
- [42] McElroy, J. L. and F. Pooler, Jr., St. Louis Dispersion Study, Volume II - Analysis, National Air Pollution Control Administration, Arlington, Virginia, Publication No. AP-53, December 1968.
- [43] Pasquill, F., Atmospheric Diffusion (D. Van Nostrand Co., Ltd., London: 1962).
- [44] Pasquill, F., "Prediction of Diffusion Over an Urban Area--Current Practice and Future Prospects," in Proceedings of Symposium on Multiple-Source Urban Diffusion Models, ed. Arthur C. Stern, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 1970.
- [45] Pooler, F., Jr., Private Communication, National Air Pollution Control Association, Raleigh, North Carolina, October 28, 1970.
- [46] Roberts, J. J., E. J. Croke and A. S. Kennedy, "An Urban Atmospheric Dispersion Model," in Proceedings of Symposium on Multiple-Source Urban Diffusion Models, ed. Arthur C. Stern, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 1970.
- [47] Savas, E. S., "Computers in Urban Air Pollution Control Systems," Socio-Economic Planning Science, 1 (1967) pp. 157-183.
- [48] Shohat, J. A. and J. D. Tamarkin, The Problem of Moments (New York: 1943).
- [49] Soyster, A. L., "Multi-Product, Multi-Period Production Scheduling via Extreme Point Properties of Linear Programming and Some Horizon Posture Properties over Time," Master of Science Thesis, Graduate School of Cornell University, Department of Operations Research, Cornell University, June 1967.
- [50] Stom, T., "Proving Strict Inequalities or Rigorously Estimating the Range of Piecewise Majorizable Functions," Dept. of Computer Sciences, Royal Institute of Technology, Stockholm, Sweden.
- [51] Sutton, O. G., Micrometeorology; A Study of Physical Processes in the Lowest Layers of the Earth's Atmosphere (McGraw-Hill, New York: 1953).
- [52] TRW Systems Group, "Proposed Implementation Plan for the Control of Particulates and Sulfur Oxides," Report SN 14838.000, prepared for the National Air Pollution Control Administration, September 1970.

- [53] TRW Systems Group, "Air Quality Implementation Program," prepared for the National Air Pollution Control Administration, November 1970.
- [54] Turner, D. B., "A Diffusion Model for an Urban Area," J. Appl. Meteorology, 3 (1964), p. 83.
- [55] Turner, D. B., Workbook of Atmospheric Dispersion Estimates, National Center for Air Pollution Control, Cincinnati, Ohio, PHS Publication No. 999-AP-26, 1967.

Entscheidungskriterien — Hilfe oder Hindernis beim Handeln?

von M. Rutsch, Karlsruhe

Man spricht heute viel von den sogenannten Entscheidungs-
wissenschaften (Decision Sciences). Zu diesem Komplex gehören

- Theorien und Techniken des Optimierens,
seine verschiedenen Varianten (z.B. dynamische
Optimierung) und die einschlägigen kombina-
torischen Hilfsmittel;
- Kontrolltheorie und Lerntheorie, sowie die
- eigentliche (statistische) Entscheidungstheorie.

Wenn es um eine Sache schlecht bestellt ist, entsteht
gewöhnlich rasch eine Wissenschaft, man denke etwa
an den expandierenden Komplex der Erziehungswissenschaften.
Der wissenschaftliche und vor allem der literarische Aus-
stoß reagiert sehr rasch und häufig überproportional auf
die neue Nachfrage. Ist daraus zu schließen, daß es wie
um das Erziehen, so auch um das Entscheiden schlecht be-
stellt ist? Einige Tatsachen sprechen dafür:

Wie beim Erziehen, so reichen auch beim Entscheiden die her-
kömmlichen Methoden, die sich unter früheren Verhältnissen
durchaus bewährt hatten, nicht mehr aus. Hinzu kommt, daß
man sich nicht mehr damit begnügen will, noch funktionie-
rende Verfahren weiter zu verwenden, ohne nachgeprüft
zu haben, warum sie funktionieren, wieweit man sich da-
rauf verlassen kann, und ob es nicht doch etwas Besseres

gibt: man ist progressiv, man experimentiert ... das gehört heute zum guten Ton. Aber auch von dieser Modeströmung abgesehen, macht sich ein tiefes und weitgehend berechtigtes Unbehagen über Entscheidungen bemerkbar, von denen wir hören. Jede Maßnahme, die sehr viele Menschen betrifft und für sie meistens eine Belastung oder Einschränkung darstellt (z.B. Steuer- oder Fahrpreiserhöhung), wird kritisiert. Die Journalisten der verschiedenen Medien fragen: War das nötig? War das richtig? Wie hören von eklatanten Fehlentscheidungen, von sinnlosen oder schädlichen Entscheidungen; von Entscheidungen, die unausgereift und unüberlegt realisiert wurden, d.h. ohne reifliche Überlegung aller Konsequenzen. Man berichtet uns von Entscheidungen, die nicht auf einwandfreie Weise zustandegekommen sind, sondern z.B. unter dem Druck von Interessengruppen. Die widersprüchlichsten Entscheidungen lösen einander in planloser Folge ab. Wir hören von Mißständen, die doch auch nichts anderes sind als das Resultat unterlassener oder hinausgeschobener Aktionen, unkoordinierter und verzögerter Entscheidungen.

Bei diesem trostlosen Stand der Dinge wäre es einfach wunderbar, wenn die Entscheidungswissenschaften den verantwortlichen Individuen oder Gremien die Arbeit des Entscheidens abnehmen und im richtigen Augenblick die richtige Entscheidung aufspüren und anzeigen würden.

Wenn das Entscheidungsproblem einen gewissen Reifegrad erreicht hat (d.h. wenn genügend viele Vorentscheidungen schon gefallen sind), besteht die Arbeit der Entscheidungsfindung darin, ein optimales Aktionsprogramm (eine optima-

le "Strategie") zu berechnen. Wir werden auf dieser Tagung Gelegenheit haben, von neuen Algorithmen und Modellen zu hören, die mit großem Scharfsinn zur Analyse, Prognose und optimalen Planung demographischer, kommunaler, ökologischer, industrieller und gesamtwirtschaftlicher Prozesse ausgedacht wurden und sicher nicht verfehlen werden, uns zu beeindrucken. Es mag sogar den Anschein haben, als gäbe es Lösungen, wo gar keine Probleme sind, oder wie Jacques Prévert so schön absurd sagt, als gäbe es "réponses à tout, questions à rien ...". Und ich brauche kaum darauf hinzuweisen, daß Optimierung, immer und überall, ein Fetisch zu werden droht und zum Teil schon geworden ist; so sagt der Chef in einer der Zeichnungen des NEW YORKER zu seinen Angestellten: "Don't just sit there - optimize something!" Trotz dieser kritischen Vorbehalte wird der Bereich, in dem OR-Techniken vernünftigerweise und mit Aussicht auf Erfolg zur Entwicklung und Lösung von Entscheidungsproblemen eingesetzt werden können und hoffentlich auch eingesetzt werden, in Zukunft immer größer werden. Darüber dürfte es nach den programmatischen Ausführungen von Herrn KONZI keinen Zweifel geben.

Problematisch ist dagegen die Hilfe der Entscheidungswissenschaften bei Aufgabenstellungen, die noch nicht soweit aufbereitet sind, daß alle wesentlichen Eingabedaten vorliegen und die optimale Aktion "nur" noch ausgerechnet zu werden braucht, und erst recht in Problemsituationen, wo eine derartige Aufbereitung einer Denaturierung gleichkäme. Kann auch da der Entscheidungstheoretiker den eigentlichen Entscheidungsträger, den für die Entscheidung Verantwortlichen, ablösen? Kann er in die Rolle des

Entscheidenden schlüpfen wie Siegfried in die Gestalt Gunthers, als dieser um Brunhild warb - oder ist auch das nur ein Mythos?

Wir wollen gleich sagen, für welche Art von Entscheidungen es nicht geht: Für die "inspirierten" Entscheidungen, die eine glückliche Eingebung des Augenblicks darstellen, für künstlerisches und allgemeiner: kreatives Schaffen. "Je ne pouvais rien faire d'après une décision prise d'avance", sagte François Mauriac in einem seiner letzten Gespräche. Ein Kunstwerk frappiert durch das Unerwartete und Einmalige; seine Entstehung kann nicht als "Auswahl unter mehreren Formen" gedacht werden - obwohl wir zuweilen seine Genese (in Vorstudien) und das tastende Suchen nach ihm verfolgen können.

Das instinktive Entscheidungsverhalten der Tiere ist stammesgeschichtlich erworben. Instinktives Entscheidungsverhalten gibt es auch beim Menschen; in manchen Situationen ist es unerlässlich, in anderen schädlich - immer aber ist es von Reflexion und Analyse bedroht, die solche instinktiven oder reaktiven Entscheidungsschemata in uns "löscht". Der Mensch muß mit seinen Entscheidungen immer wieder ganz vorn, am Nullpunkt anfangen; das ist ein Privileg und zugleich eine große Belastung. Daher ist der Mensch im allgemeinen froh, wenn er Entscheidungen los wird. Entscheidungen machen, wie wir alle aus eigener Erfahrung wissen, müde und reizbar. Wir können sehr viel mehr Wahrnehmungen aufnehmen als Entscheidungen treffen - ich glaube 200 bits/min gegenüber 10 - 50 bits/min - und wir ermüden bei 50 bits Entscheidungen pro Minute sehr schnell. Das an sich schon anstrengende Geschäft des Entscheidens wird

häufig noch dadurch erschwert, daß der Entscheidungsträger sich in seiner Rolle nicht wohlfühlt. Psychologen sehen es geradezu als Indiz für schlechte Anpassung in einer beruflichen Stellung an, wenn jemand Angst hat, selbst kleinere Entscheidungen zu treffen. Häufig sind pathologische Störungen des Entscheidungsverhaltens zu beobachten, z.B. die Unfähigkeit, die der Position des Betreffenden angemessenen Entscheidungen zu treffen, oder die Neigung, solche Entscheidungen durch eine übergewissenhafte Analyse endlos hinauszuschieben und damit zu umgehen. In dem Büchlein "The Peter Principle" von L. J. PETER und R. HULL findet man etliche fingierte, aber zutiefst realistische Krankheitsbilder zur Illustration gehemmten oder blockierten Entscheidungsvermögens.

Aus all diesen Gründen sucht der Mensch einen Freund, Beichtvater, Advokaten, Astrologen, Mathematiker oder sonst eine "Vatergestalt", die ihm seine Entscheidungen abnimmt. Ein Kollektiv delegiert die Entscheidungsbefugnis an ein Führungsgremium oder an eine Einzelperson, z.B. an eine Persönlichkeit, die durch die Kühnheit ihrer politischen Konzeption imponiert, oder an jemanden, der "Glück" hat oder im Rufe steht, Glück zu haben - ein wichtiger Gesichtspunkt im Altertum bei der Wahl eines Feldherrn für eine kriegerische Expedition. Durch solche Übertragung der Entscheidungsbefugnis hofft man, sich gegen Fehlentscheidungen abzusichern, und manchmal hat der Berater oder neue Entscheidungsträger tatsächlich bessere Kenntnisse, größere Erfahrung oder größeres Geschick in der Lenkung komplexer und langfristiger Unternehmungen.

Der Stellvertreter, dem wir das Entscheiden auftragen, ist

uns natürlich "verantwortlich" und in gewissem Maße haftbar für die Schäden, die durch seine Amtsführung evtl. entstehen. Dadurch versucht man, ihn psychologisch und mit seiner sozialen und materiellen Existenz an den Erfolg des Unternehmens zu binden. Das ist ein sehr komplexer, um nicht zu sagen zwielichtiger Vertrag, eine Art Seelenhandel oder Pakt mit dem Teufel (Wer ist der Teufel?). Es kommt vor, daß die Beteiligten ein unverschuldetes Fiasko mit ihrer Existenz bezahlen - beispielsweise der Verantwortliche, auch wenn er weder (absichtlich oder fahrlässig) den Zielvorstellungen des Kollektivs zuwider entscheiden hat, noch riskanter vorging als dem Kollektiv akzeptabel erschien. Andererseits kann ein Einzelner durch seine persönliche, selbst unbegrenzte Haftung den durch seine Entscheidung entstandenen Schaden eines großen Kollektivs nicht reparieren. Zu dieser Problematik der Entscheidungsdelegation kommt die ebenfalls grundsätzliche Schwierigkeit hinzu, dem neuen Entscheidungsträger die richtigen Maßstäbe für seine Entscheidungstätigkeit mitzugeben: Welche Wert- und Zielvorstellungen sollen zugrundegelegt werden? Welche Haupt- und Nebenwirkungen sind (getrennt oder vereint) zu veranschlagen? Welches Risiko darf man äußerstenfalls eingehen?

Das sind, bei Entscheidungen, die ein Kollektiv betreffen, außerordentlich brisante wirtschaftliche oder gesellschaftliche Vor-Entscheidungen. Sie müssen von den Betroffenen im wesentlichen selbst entschieden werden, und kein Entscheidungswissenschaftler kann ihnen diese Arbeit abnehmen. Die Aktionäre einer Versicherungsgesellschaft vertrauen die Führung der Geschäfte dem Vorstand und dem Aufsichtsrat an, deren Mitglieder auf Grund ihrer Fähigkeiten, Erfahrungen, Zuverlässigkeit und wegen ihres erwiesenen oder erwarteten maßvollen Risikoverhaltens ausgewählt wurden. Dieser unvermeidbaren, durchaus üblichen und in den meisten Fällen gerechtfertigten Entscheidungsdelegation schließt

sich möglicherweise eine zweite an: Die genannten Gremien versuchen ihre Entscheidungsbefugnis weiterzudelegieren an einen Versicherungsmathematiker, wenn sie z. B. vor der Frage stehen: In welcher Höhe soll sich die Gesellschaft rückversichern, wenn bei ihr ein 10 Millionen ₯ teures Schiff gegen eine jährliche Prämie von 300000 ₯ versichert ist? KARL BORCH, bei dem man dieses Beispiel mit weiteren, amüsanten Einzelheiten und sarkastischen Bemerkungen nachlesen kann (The Economics of Uncertainty, 2.13 ff.), bezweifelt, daß ein noch so gelehrtes statistisches Gutachten mehr als numerische Details zur Beantwortung der Frage beisteuern kann; das Entscheidungsproblem ist schon auf seine einfachste, nicht weiter reduzierbare Form gebracht und muß von den Leitungsgremien der Gesellschaft in dieser Form bewältigt werden.

Wann immer wir eine Entscheidung delegieren, vertrauen wir einem z.T. unbekannten Mechanismus die Entscheidungsfindung und damit unser Schicksal an und treffen damit die folgeschwerste Entscheidung. Das gilt auch für die Hilfe, die wir von Entscheidungswissenschaften, speziell Entscheidungstheorie und Entscheidungskriterien erwarten: ohne Kenntnis der dort verwendeten Prinzipien, ohne Prüfung, ob wir die vom Entscheidungsmodell benötigten Daten mit der notwendigen Präzision bereitstellen können, wird die erhoffte Hilfe u.U. eher eine Gefahr - und jedenfalls ein unkalkuliertes Risiko - darstellen.

Nehmen wir an, es sei uns gelungen, alle diese vielfältigen und vielschichtigen Bedenken auszuräumen! Dem Entscheidungstheoretiker sei es also gestattet, sich an die Stelle des ursprünglichen Entscheidungsträgers zu versetzen. Dann bleibt noch eine grundsätzliche und beunruhigende Schwierig-

keit: die Ungewißheit. Gemeint ist die Ungewißheit darüber, was bei jeder der ins Auge gefaßten Handlungsalternativen herauskommt. Welches sind die erkenntnislogischen Wurzeln dieser Ungewißheit? Ihr Entstehungsgrund ist in erster Linie darin zu sehen, daß die Handlungsalternativen durch Aussagen

$$a_1, a_2, \dots, a_n$$

beschrieben werden und die Umweltzustände ebenfalls durch Aussagen

$$b_1, b_2, \dots, b_m$$

beschrieben werden. Im Rahmen einer "Theorie", die unser einschlägiges Wissen verkörpert, können wir Implikationen, etwa

$$a_i \wedge b_j \Rightarrow e_{ij},$$

ableiten. Die Aussage e_{ij} beschreibt die gemeinsame Konsequenz der Handlung a_i und des Umweltzustandes b_j nach unserem besten Wissen. Das heißt: Wenn ich das freigewählte a_i realisiere (= die Aussage a_i "wahr mache") und weiß, daß b_j zutrifft, dann wird die Konsequenz e_{ij} Wirklichkeit. Wenn ich aber z.B. nur weiß, daß b_j oder b_k zutrifft, dann ist auch die Handlungskonsequenz nur ungenau bekannt:

$$a_i \wedge (b_j \vee b_k) \Rightarrow e_{ij} \vee e_{ik}.$$

Natürlich sind auch die ursprünglichen Konsequenzaussagen e_{ij} nicht absolut präzise und erschöpfend; aber selbst wenn wir

annehmen, daß sie alle Angaben in hinreichender Präzision enthalten, die den von der Entscheidung Betroffenen interessieren dann kann doch bereits die ungenauere Aussage

$$e_{ij} \vee e_{ik}$$

eine beängstigende Ungewißheit über den Ausgang der Aktion a_i darstellen. Daran knüpfen sich eine Reihe von Aufgaben und Problemstellungen. Wir wollen einige von ihnen kurz skizzieren.

Das Ordnen der Konsequenzen gemäß unseren Präferenzen bereitet Schwierigkeiten, da es sich bei ihnen prinzipiell um ungefähre Beschreibungen

$$e_{ij} \text{ oder } e_{ij} \vee e_{ik} \text{ usw.}$$

von Sachverhalten handelt, die unser Wohlergehen berühren. Man kann sich diese ungefähren Ergebnisse etwa als Teilmengen (Punkthaufen oder Intervalle) auf einer Skala vorstellen, nicht als präzise Punkte darauf. Die Schwierigkeiten werden verschärft, wenn wir die Aktion a_i nicht genau festlegen können oder b_j nicht genau kennen (s.o.). Die Schwierigkeiten werden gemildert, wenn wir die Umweltaussage b_j präzisieren können, oder wenn uns eine bessere Theorie präzisere Implikate e_{ij} zu den Prämissen a_i und b_j liefert.

Die - nach Möglichkeit totale - Präferenzordnung der e_{ij} dient dem Zwecke, die Aktionen zu ordnen, die eine oder mehrere dieser Konsequenzen hervorbringen können, und damit

letzten Endes zur Auswahl einer optimalen Aktion. Selbst wenn die e_{ij} totalgeordnet sind (d.h. je zwei von ihnen sind präferenzmäßig vergleichbar), wir aber nicht wissen, welches b_j zutrifft, können wir deswegen trotzdem die Aktionen nicht ohne weiteres auch total ordnen, weil ja jeder Aktion ein ganzes Bündel von Konsequenzen entspricht. Die Entscheidungskriterien haben gerade den Zweck, auch unter solchen Umständen aus der (Präferenz-)Ordnung der Konsequenzen eine (Präferenz-)Ordnung zwischen den Aktionen herzuleiten. Wir müssen also Entscheidungskriterien entwickeln für den Fall, daß die Konsequenzen der einzelnen Aktionen nicht eindeutig bekannt sind. Beispielsweise sei der wahre Umweltzustand nur in Form der ungenauen Aussage \bar{b} bekannt und von den zwei Handlungen a_1, a_2 wisse man,

bei \bar{b} kann a_1 zu e_1 oder auch zu e'_1 führen,
 a_2 zu e_2 oder auch zu e'_2 .

Man kann \bar{b} formal in vier feinere Zustände aufspalten und erhält

	b_1	b_2	b_3	b_4
a_1	e_1	e'_1	e_1	e'_1
a_2	e_2	e_2	e'_2	e'_2

Dieser schöne Trick nützt garnichts für die Entscheidung zwischen a_1 und a_2 , außer in speziellen Situationen wie den folgenden:

1. Jede bei \bar{b} mögliche Konsequenz von a_1 ist besser als jede Konsequenz von a_2 : dann ziehen wir a_1 der Aktion a_2 vor. Dieses selbstverständliche Vorgehen läuft darauf hinaus, daß wir zwischen den Teilmengen der Menge E aller möglichen Konsequenzen eine Ordnung \succ_1 einführen durch:

$$\{e_1, e_1', \dots\} \succ_1 \{e_2, e_2', \dots\}$$

falls jedes e_1, e_1', \dots besser ist als jedes e_2, e_2', \dots . Diese partielle Ordnung auf $\mathcal{P}(E)$ überträgt sich in offensichtlicher Weise auf die Menge A der Aktionen und kann eventuell sogar, wie im obigen speziellen Fall, auf der Teilmenge der zur Entscheidung stehenden Aktionen (oben: a_1 und a_2) eine totale Ordnung induzieren.

2. Wir sind bereit, uns nach dem Maximin-Prinzip zu entscheiden. Das bedeutet: eine Aktion ist uns um so lieber, je weniger schlechte Konsequenzen sie hervorbringen kann. Dieser Einstellung entspricht folgende Ordnung \succ_2 auf $\mathcal{P}(E)$:

$$\min \{e_1, e_1', \dots\} \succ \min \{e_2, e_2', \dots\},$$

wo \succ die Präferenzordnung auf E bezeichnet, Ist \succ total, so gilt dasselbe von \succ_2 . Die Ordnungsrelation \succ_2 enthält \succ_1 .

3. Wir sind in der Lage, mehr über die Möglichkeit bzw. Unmöglichkeit der feineren Zustände b_1, b_2, \dots herauszufinden. Wir mögen beispielsweise feststellen, daß

$$e_1 \text{ und } e_2 \text{ unvereinbar}$$

sind und, daß

$$e_1 \Rightarrow a_2 \text{ und } e_2 \Rightarrow a_1.$$

Dann kann es keinen Feinzustand b_1 geben (man beachte, daß die Unvereinbarkeit von e_1 und e_2 dazu nicht ausreicht). Nach Streichung der zu b_1 gehörigen Spalte in der obigen Ergebnismatrix mag sich vielleicht Dominanz (im Sinne von \succ_1 , siehe Nr. 1) zwischen a_1 und a_2 herausstellen, etwa wenn

$$e'_1 \succ e_2 \succ e_1 \succ e'_2.$$

Die Streichung von Feinzuständen nach Nr. 3 zerstört nichts an der Relation \succ_1 der Nr. 1, sondern vergrößert sie höchstens. Dagegen kann die Relation \succ_2 , die schon maximal ist, u.U. in ihr Gegenteil verkehrt werden. Beides ist aber nur möglich, wenn eine Konsequenz, z.B. e_2 , als Ergebnis von a_2 völlig ausfällt, wenn also im obigen Beispiel b_1 und b_2 gestrichen werden - das entspricht einer Einengung des Ergebnisses von a_2 und \bar{b} von $\{e_2, e'_2\}$ auf $\{e'_2\}$

Die letzte Überlegung ist wichtig, wenn wir es nicht mit einem Grobzustand \bar{b} zu tun haben, sondern mit mehreren Grobzuständen $\bar{b}_1, \bar{b}_2, \dots$. Es liegt dann eine "ungenau" Ergebnismatrix

	\bar{b}_1	\bar{b}_2	...
a_1	E_{11}	E_{12}	...
a_2	E_{21}	E_{22}	...
\vdots	\vdots	\vdots	

vor mit Ergebnissen E_{ij} , die Teilmengen von E sind. Als Prä-

ferenzordnung zwischen den E_{ij} (d.h. auf $\mathcal{P}(E)$) können wir \succ_1 oder \succ_2 zugrundelegen. Es werde angenommen, daß über der Menge der Grobzustände völlige Ungewißheit herrsche. Ähnlich wie vorhin können wir dann wieder in speziellen Situationen zu einer Entscheidung finden:

1. Wir stellen möglicherweise Dominanz von a_1 über alle anderen Aktionen a_i ($i > 1$) fest, d.h.

$$\forall i > 1 \quad \forall j: E_{1j} \succ_1 E_{ij}.$$

Dann werden wir selbstverständlich a_1 zu realisieren suchen. Wenn wir uns jedes \bar{B}_j als eine Menge von Feinzuständen vorstellen, dann ist die eben angesprochene Dominanz gleichbedeutend mit Dominanz bzw. der Menge $\bigcup \bar{B}_j$ von Feinzuständen.

2. Wir entscheiden uns nach dem Maximin-Prinzip. Dann kann die obige Ergebnismatrix ersetzt werden durch folgende einfachere Matrix:

	\bar{B}_1	\bar{B}_2
a_1	$\min E_{11}$	$\min E_{12}$
a_2	$\min E_{21}$	$\min E_{22}$
	\vdots	\vdots	

und wir können auf sie das Maximinkriterium zur Entscheidungsfindung anwenden. Ein interessantes Mischkriterium ergibt sich übrigens, wenn man stattdessen das Dominanzprinzip auf die letzte Ergebnismatrix anwendet. Dieselbe Ergebnis-

matrix kann ferner auch im Risikofall (s.u.) zugrundegelegt werden.

Die eben besprochene Entscheidungslage - völlige Ungewißheit über einer Menge von Grobzuständen $\overline{B}_1, \overline{B}_2, \dots$ - kann man sich so entstanden denken, daß zunächst ein ganz grober Zustand \overline{B} vorlag. \overline{B} wurde nach dem früher erwähnten Trick in Feinzustände zerlegt, von denen einige nach Nr. 3 gestrichen werden konnten. Mehrere Feinzustände, deren Ergebnisse bei a_1, a_2, \dots in beliebigen Kombinationen vorkommen können (d.h. keine solche Ergebniskombination wurde gestrichen), werden schließlich zu einem \overline{B}_j zusammengefaßt.

Bei derartigen Umformungen und überhaupt beim Vergleich von Entscheidungsproblemen bemerken wir die allgemeine Wirksamkeit folgender Komplementarität: Je feiner wir die Fallunterscheidungen

$$a_1, a_2, \dots \quad \text{und} \quad b_1, b_2, \dots$$

machen, um so präziser werden natürlich die e_{ij} , aber um so schwerer ist es auch, das wahre b_j herauszufinden oder präzisere Wahrscheinlichkeitsaussagen darüber zu beschaffen. Betrachten wir etwa, als Gegenstück zur völligen Ungewißheit, den Risikofall für Grobzustände! Gegeben seien also präzise numerische Wahrscheinlichkeiten $\overline{p}_1, \overline{p}_2, \dots$ für die Grobzustände $\overline{B}_1, \overline{B}_2, \dots$. Spalten wir jedes

$$\overline{B}_j = b_j^1 + b_j^2 + \dots$$

in feinere Zustände auf, so tritt an die Stelle der eindeutigen Wahrscheinlichkeitsverteilung über $\overline{B}_1, \overline{B}_2, \dots$ eine Menge von Wahrscheinlichkeitsverteilungen

$$b_j^k \mapsto p_j^p,$$

über der wieder völlige Ungewißheit herrscht. Es gibt übrigens zwei Auffassungen über diese Menge von Wahrscheinlichkeitsverteilungen: man kann alle Verteilungen mit $\sum_k p_j^k = \bar{p}_j$ betrachten, oder alle Verteilungen mit $p_j^k = \bar{p}_j$ für jeweils ein $k = k(j)$, $j = 1, 2, \dots$.

Die völlige Ungewißheit kommt also durch die Hintertür wieder herein und muß durch ein Kriterium wie das Maximin-kriterium bewältigt werden. Gleichbedeutend ist die Anwendung des Bayesschen Kriteriums auf die Entscheidungsmatrix

	\bar{p}_1	\bar{p}_2
	\bar{b}_1	\bar{b}_2
a_1	$\min E_{11}$	$\min E_{12}$
a_2	$\min E_{21}$	$\min E_{22}$
\vdots	\vdots	\vdots	

unter Verwendung einer erwartungstreuen Nutzenfunktion auf E . Das ist der Sinn des von G. MENGES so genannten Adaptationskriteriums oder des SCHNEEWEISS-Kriteriums.

Wie wir jetzt gesehen haben, kann sich eine zunächst vorhandene präzise Wahrscheinlichkeitsverteilung zu einer (begrenzten) Ungewißheitssituation "verwischen" oder "verwackeln", wenn man nämlich die Zustände verfeinert. Setzt man aber in der statistischen Entscheidungstheorie

nicht voraus, daß auf dem Raum auch der feineren Zustände eine wohlbestimmte Wahrscheinlichkeitsverteilung vorliegt, oder verspricht man nicht gar, eine solche zu beschaffen? In der Tat erwecken einige Entscheidungstheorien diesen Anschein, beispielsweise diejenige von LEONARD J. SAVAGE, die in verfeinerter Form auch auf dieser Tagung zur Sprache kommen wird. Der Entscheidungstheoretiker legt dem (fiktiven) Individuum, dem er bei der Entscheidung helfen will, Lotterien zur Auswahl vor und destilliert aus dem (hoffentlich) konsistenten Wahlverhalten des Individuums dessen persönliche Wahrscheinlichkeitsverteilung. Schauen wir uns diesen Prozeß einmal etwas näher an, um herauszufinden, ob er tatsächlich zu einer präzisen Wahrscheinlichkeitsverteilung führt!

Wir benutzen zu diesem Zweck ein sehr einfaches Modell, dessen zentrale Figur ein Zufallsexperiment mit zwei möglichen Ausgängen ist. An jeden Ausgang können wir ein Ergebnis e_1 , e_2 usw. knüpfen. Mit

$$e_1 \circ e_2$$

bezeichnen wir die Lotterie, bei der das betreffende Experiment einmal durchgeführt wird und beim Eintreffen des ersten bzw. zweiten Ausganges e_1 bzw. e_2 als Preis ausgehändigt wird. Wir verwenden nur zwei Ergebnisse, ein "gutes", e , und ein "schlechtes", 0 :

$$0 \prec e.$$

Daraus können wir aber mit Hilfe der Verknüpfung \circ unendlich

viele zusammengesetzte Lotterien bilden, z.B.

$$0 \circ 0, e \circ e, 0 \circ e, e \circ 0, (0 \circ e) \circ e \text{ usw.}$$

Auf diese Weise entsteht eine freie Wortalgebra mit der Verknüpfung \circ und dem Alphabet $\{0, e\}$, auf der eine Präferenzordnung definiert sei. Letztere entspricht dem Wahlverhalten des Individuums gegenüber den Lotterien. Die geforderte Konsistenz dieses Wahlverhaltens drückt sich in einigen Relationen aus wie

$$0 \circ 0 \sim 0$$

oder

$$0 \circ e \prec e \circ 0 \rightarrow (0 \circ e) \circ e \prec (e \circ 0) \circ e$$

oder der PFANZAGLschen Bisymmetrie und noch anderen Relationen. Wir illustrieren diese Konsistenzforderungen an den ersten Schritten des Auswahlprozesses und zeigen, wie sie eine sukzessive numerische Implementation der Wahrscheinlichkeit ermöglichen.

Zunächst hat sich unser Individuum zu entscheiden für

$$0 \circ e \succ e \circ 0 \quad \text{oder} \quad 0 \circ e \prec e \circ 0.$$

Das ist gleichwertig zur Frage: Dünkt ihm der zweite Ausgang wahrscheinlicher als der erste, oder ist es umgekehrt? Und für die zunächst noch fraglichen numerischen Wahrscheinlichkeiten p des ersten Ausgangs und $q = 1-p$ des zweiten

Ausgangs bedeutet es die Festlegung auf

$$p < q, \text{ d.h. } p < \frac{1}{2} \quad \text{oder} \quad p > q, \text{ d.h. } p > \frac{1}{2}.$$

Wir nehmen an, das Individuum habe sich für die zweite Möglichkeit entschieden. Dadurch hat es für die dreigliedrigen Lotterien auch schon entschieden, daß

$$0 \circ 0 \circ 0 < 0 \circ (0 \circ e) < 0 \circ (e \circ 0) < (0 \circ e) \circ 0 \\ < (e \circ 0) \circ 0.$$

Frei entscheiden kann sich das Individuum aber noch für (oder gegen)

$$(e \circ 0) \circ 0 \succ (0 \circ 0) \circ e.$$

Hat es sich dafür entschieden, dann verfeinert bzw. erweitert sich die obige Präferenzkette zu

$$\dots (0 \circ e) \circ 0 < 0 \circ e \sim (0 \circ 0) \circ e < (e \circ 0) \circ 0.$$

Jetzt ist nur noch offen, ob man die Lotterie $(0 \circ e) \circ e$ zwischen die letzten beiden Glieder der Präferenzkette stellen will oder rechts davon. Wenn auch das geklärt ist, sind die dreigliedrigen Lotterien alle in eine totale Ordnung gebracht. Was bedeuten die letzten beiden Entscheidungen für die numerischen Wahrscheinlichkeiten p und q ?

$(e \circ 0) \circ 0 \succ (0 \circ e) \circ e$ bedeutet $p^2 > pq + q$ und somit $p > \frac{1}{2}\sqrt{2}$; $(e \circ 0) \circ 0 \succ (0 \circ 0) \circ e$ folgt notwendig aus der vorige

Präferenzentscheidung.

$(0 \circ e) \circ e \succ (e \circ 0) \circ 0 \succ (0 \circ 0) \circ e$ bedeutet

$pq + q > p^2 > q$, d.h. $\frac{1}{2} \sqrt{2} > p > \frac{1}{2} (\sqrt{5}-1)$.

Schon diese allerersten Schritte des Präferenzentscheidungsprozesses lassen erkennen, daß die unbekannten Wahrscheinlichkeiten (wir hatten es oben nur mit einer zu tun, nämlich mit p) mit wachsender Genauigkeit in immer kürzere Intervalle eingeschlossen werden - daß sie aber nie mit endgültiger Präzision festliegen. Wahrscheinlichkeitsintervalle statt exakter, punktförmiger Wahrscheinlichkeiten wurden auch von B. O. KOOPMAN und von I.J. GOOD untersucht, die mit "oberen" und "unteren" Wahrscheinlichkeiten arbeiten. Ebenso wie mit ungenauen Ergebnisangaben muß die Entscheidungstheorie auch mit solchen unscharfen Wahrscheinlichkeitswerten umzugehen lernen. Solange sie das nicht kann, besteht die Gefahr, daß sie die groben empirischen Informationen mit vorgetäuschter Präzision zu einer "optimalen" Entscheidung verarbeitet.

Kommunale Anwendungen

Digitale Simulation der räumlichen Stadtentwicklung von J. Meise und M. Wegener, Frankfurt a. M.

1. Simulation sozioökonomischer Systeme

Mathematische Simulation ist eine Form des wissenschaftlichen Experiments, bei der, anstatt mit der Wirklichkeit selbst, mit einem numerisch-logischen Modell der Wirklichkeit experimentiert wird.

Durch ihren experimentellen Ansatz unterscheidet sich mathematische Simulation von analytischen Lösungsverfahren mathematischer Systeme, mit denen eine Variable des Systems aus den übrigen Variablen allgemein errechnet wird. Simulation bedeutet dagegen die Ermittlung des Systemzustands für eine spezielle Konstellation von Werten der Systemvariablen. Ebenso unterscheidet sich mathematische Simulation von Optimierungsverfahren, mit denen auf analytischem Wege diejenigen Werte der Systemvariablen festgestellt werden, die eine Zielfunktion unter Berücksichtigung der einschränkenden Bedingungen des Systems am besten erfüllen. Die Ermittlung einer optimalen Lösung ist in der Regel nicht das Hauptziel der Simulation. Selbstverständlich dienen auch Simulationsverfahren letzten Endes der "Optimierung" des untersuchten Systems. Wesentliches Ziel der Simulation ist jedoch die Gewinnung von Erkenntnissen über das Verhalten des beobachteten Systems unter unterschiedlichen Bedingungen.

Mathematische Simulationsverfahren gewinnen wachsende Bedeutung für die Erforschung und Planung sozioökonomischer Systeme. Die Organisation der menschlichen Gesellschaft als Ganzes

und ihrer Teilbereiche besitzt Eigenschaften eines komplexen dynamischen Systems. Viele ihrer Wirkungszusammenhänge sind zirkulär und können als positive oder negative Rückkopplungsschleifen interpretiert werden. Es ist möglich, Teilbereiche des sozialen und ökonomischen Gesamtsystems in Form von Differential- oder Differenzengleichungen darzustellen.

Allerdings entziehen sich solche Gleichungssysteme häufig einer analytischen Lösung: die meisten beobachteten Abhängigkeiten sind prinzipiell nichtlinear (Wachstumsvorgänge, Verhaltenswahrscheinlichkeiten, Verteilungs- und Kostenfunktionen); viele von ihnen sind sogar nicht glatt oder nicht stetig, haben dichotomischen oder Stufencharakter. Zudem kann bei nur geringer sachlicher, räumlicher und zeitlicher Differenzierung des Modells die Zahl der Gleichungen des Systems sehr groß werden.

Simulationsverfahren dagegen erlauben es, auch umfangreiche Systeme mit beliebigen numerischen und logischen Verknüpfungen mathematisch einfach und anschaulich in ihrer zeitlichen Entwicklung darzustellen.

Aber nicht nur deshalb finden mathematische Simulationsverfahren wachsende Anwendung auch im Bereich der sozioökonomischen Planung. Der experimentelle Charakter der Simulation entspricht in besonderer Weise dem iterativ ablaufenden Entscheidungsprozeß bei der Gestaltung der menschlichen Umwelt. Dieser unterscheidet sich, trotz prinzipieller Ähnlichkeit der Ziel-Mittel-Relation, nicht unwesentlich vom Ablauf überwiegend technisch-wirtschaftlicher Planungen: Während dort in der Regel ein einziges, eindeutig und operational formuliertes Planungsziel vorliegt, muß bei Planungen im sozioökonomischen Bereich typischerweise von mehreren, nicht miteinander vereinbaren und

nicht einmal in sich selbst widerspruchsfreien, komplexen Zielbündelungen ausgegangen werden, die von verschiedenen Gruppen der Gesellschaft mit unterschiedlichem Nachdruck und häufig auf unterschiedlichen Ebenen des Problemverständnisses vertreten werden. Der Weg vom Erkennen eines Planungsproblems bis zur schließlichen Entscheidung für eine konkrete Maßnahme zu seiner Lösung ist hier ein mühsamer und langwieriger Prozeß, in dem nicht nur die Lösungen, sondern auch die Ziele mehrfach korrigiert werden müssen, bis ein befriedigender Ausgleich zwischen den divergierenden Zielen und den zur Verfügung stehenden Mitteln gefunden worden ist.

Experimente mit Simulationsverfahren können diesen Prozeß der schrittweisen Annäherung an eine Problemlösung in hervorragender Weise unterstützen. Als besonderer Vorteil erweist sich dabei, daß die Arbeit mit einem Simulationsmodell relativ voraussetzungslos, d.h. mit geringem Informationsstand über das Planungsproblem selbst, über die Konstellation der Ziele und ihre möglichen Konflikte begonnen werden kann. Optimierungsverfahren erfordern dagegen als ersten Schritt die Formulierung einer hier notwendigerweise mehrdimensionalen Zielfunktion mit all ihrer grundsätzlichen Problematik. Bei der Simulation erfolgen Bewertung und Auswahl mit gutem Grund außerhalb des eigentlichen Modells. Stattdessen soll die Arbeit mit dem Modell einen Lernprozeß über die Zusammenhänge des untersuchten Systems, über die Folgen und Wechselwirkungen von Planungseingriffen ingangsetzen, der es ermöglicht, iterativ zu immer "besseren" Lösungen vorzustoßen.

In jüngster Zeit mehren sich die Bemühungen, das Konzept der mathematischen Simulation auch auf die Probleme der langfristigen Stadtentwicklung zu übertragen. Die Stadt wird dabei verstanden als ein komplexes, dynamisches System sozialer, ökonomischer und technischer Beziehungen, das sich in seiner räumlichen und zeitlichen Dimension verändert. Die Veränderungen des Systems können in begrenztem Maße durch Planungsein-

griffe beeinflußt oder gesteuert werden. Die langfristige Planung dieser Eingriffe ist Stadtentwicklungsplanung.

2. Das POLIS-Modell

2.1 Zielsetzung

Seit 1969 wird am Battelle-Institut an einem Modell als Entscheidungshilfe für die Stadtentwicklungsplanung gearbeitet. Das Modell setzt sich aus drei Teilmodellen zusammen, die stufenweise aufeinander aufbauen (Abb. 1):

- Das erste Teilmodell (Prognosemodell) soll die Möglichkeiten und Grenzen des Bevölkerungs- und Wirtschaftswachstums für die Stadt als Ganzes prognostizieren.
- Das zweite Teilmodell (Simulationsmodell) soll die Auswirkungen alternativer Planungsvarianten der räumlichen Stadtentwicklung erkennbar machen.
- Das dritte Teilmodell (Bewertungsmodell) soll durch die Bereitstellung eines überschaubaren Bewertungsverfahrens die Rationalität der Entscheidung zwischen den Planungsvarianten erhöhen helfen.

Von den drei Teilmodellen liegt jetzt das zweite Teilmodell, das Simulationsmodell, in anwendbarer Form vor. Die erste Erprobung des Modells mit Daten einer konkreten Stadt ist inzwischen abgeschlossen. Hierzu stellte das Stadtplanungsamt der Stadt Köln Daten über die Bevölkerungs-, Wirtschafts-, Bebauungs- und Verkehrsstruktur Kölns zur Verfügung. Die weitere Erprobung und Anwendung des Modells für Planungsaufgaben Kölns erfolgt, in enger Abstimmung mit dem Stadtplanungsamt, im Rahmen eines Forschungsauftrags des Bundesministeriums für Städtebau und Wohnungswesen.

Das Simulationsmodell POLIS ist eine Entscheidungshilfe bei der langfristigen Planung der räumlichen Stadtentwicklung. Es macht die vielschichtigen Auswirkungen alternativer Kombina-

tionen von Planungsmaßnahmen auf den Gebieten Flächennutzung und Verkehr in ihrer gegenseitigen Abhängigkeit unter Berücksichtigung des Zeitablaufs erkennbar und liefert Basisinformationen für die mittel- bis langfristige Investitionsplanung der Stadt.

Jede Stadt steht laufend vor grundlegenden Entscheidungen zur räumlichen Stadtentwicklung. Wo soll ein neues Wohngebiet angelegt werden? Soll man in der City weitere Hochhäuser zulassen? Sollte man mehr für den Straßenbau ausgeben? Oder den U-Bahn-Bau beschleunigen? Oder Tiefgaragen in der City bauen? Die Diskussion über solche Fragen wird nicht nur durch widerstreitende Interessen verdunkelt, sondern auch durch den Umstand, daß niemand verlässliche Angaben über die wahrscheinlichen Folgen der möglichen Handlungsalternativen machen kann. Der Vergleich zwischen ihnen, eine Bewertung ihrer Vor- und Nachteile nach rationalen Kriterien ist unmöglich. Die Folge ist bestenfalls die Lähmung aller Entscheidungen, im schlimmeren Falle die Entscheidung für einen schlecht geprüften Plan.

Hier setzt das POLIS-Modell ein. Es ermöglicht, schnell und mit geringem Aufwand eine Vielzahl von Handlungsalternativen der Stadtverwaltung experimentell durchzuspielen. Es trägt durch eine differenzierte Darlegung der wichtigsten technischen, ökonomischen und sozialen Konsequenzen der einzelnen Entscheidungsalternativen dazu bei, diejenigen Kombinationen von Planungsmaßnahmen auszuwählen, welche den Zielvorstellungen und finanziellen Möglichkeiten der Gemeinde am besten entspricht.

2.2 Die Stadt als System

Die Teilsysteme Flächennutzung und Verkehr und die Beziehungen zwischen ihnen bilden zusammen das räumlich-zeitliche System Stadt. Es erhält seine räumliche Dimension durch die Einteilung des Stadtgebiets und seines Umlands in eine Anzahl von Teilzonen, seine zeitliche dadurch, daß nicht nur ein einziger

Systemzustand untersucht wird, sondern die Veränderung des Systems über mehrere Zeitschritte (Perioden) hinweg.

Die Beziehungen zwischen Flächennutzung und Verkehrsaufkommen sind seit langem bekannt und bilden die Grundlage jeder Verkehrsprognose. Aus prognostizierten Flächennutzungsdaten der Teilzonen des Untersuchungsgebiets wird auf die zukünftigen Verkehrsströme zwischen ihnen geschlossen. Der Ausbau des Verkehrssystems erfolgt entsprechend der prognostizierten Verkehrsnachfrage. Weit weniger erforscht sind die Einflüsse, die vom Verkehrssystem, also dem Verkehrsangebot her, auf die Flächennutzung einwirken. Dabei gehören diese Einflüsse zu den Hauptantriebskräften der räumlichen Stadtentwicklung. Im POLIS-Modell werden sie mit Hilfe der Variablen "Zugänglichkeit" erfaßt. "Zugänglichkeit" ist ein Maß für die Lagegunst einer Zone in bezug zu den Aktivitäten in allen Zonen unter Berücksichtigung der vorhandenen Verkehrsverbindungen. Eine Verbesserung der Erschließung einer Teilzone erhöht ihre Zugänglichkeit und damit ihre Attraktivität für die Errichtung von Wohnungen und Betrieben, und, sofern sie über Baulandreserven verfügt, wird sie, ceteris paribus, mehr Neubautätigkeit auf sich ziehen als andere, weniger gut erschlossene Zonen. Als Folge erhöht sich das Verkehrsaufkommen von und zu dieser Zone, das Verkehrssystem muß weiter ausgebaut werden, und so weiter ... (Abb. 2).

Dieser Kreislauf der "Stadtentwicklung" findet sein Ende an einer Grenze, die durch die technischen Möglichkeiten bzw. die Bereitschaft der Gesellschaft, diese Möglichkeiten zu erproben, gesetzt wird. Die meisten großen Städte haben diese Grenze heute bereits erreicht: Das Verkehrsaufkommen übersteigt die Kapazität der Verkehrswege, die Zugänglichkeit der am stärksten belasteten Teilzone sinkt trotz verstärkter Verkehrsinvestitionen. Gleichzeitig beansprucht gerade der Ausbau der Verkehrsanlagen immer größere Anteile des knappen Bodens, der damit anderen Nutzungen entzogen wird.

2.3 Logik des Modells

Die genannten Teilsysteme und Zusammenhänge bilden das logische Grundgerüst des POLIS-Modells. Der Zustand der einzelnen Teilzonen des Untersuchungsgebiets (interne Zonen) wird durch Datengruppen wie "Einwohner nach Altersgruppen", "Beschäftigte nach Wirtschaftsbereichen", "Wohnungen nach Baualter und Bauzustand" und "Flächen nach Nutzungsarten", der Einfluß der Umlandzonen (externe Zonen) durch ihre Einwohner, Arbeitsplätze und Erholungsgebiete im Modell repräsentiert. Das Verkehrssystem wird durch Streckendaten der wichtigsten Teilstrecken der beiden konkurrierenden Verkehrsarten, des öffentlichen Personennahverkehrs (OPNV) und des Individualverkehrs auf der Straße (PKW) im Modell dargestellt. Die Streckendaten enthalten u.a. Angaben über Streckenart, Streckenlänge, Reisezeit, Kapazität, verkehrende Linien und Zugfolge.

Diese Inventardaten werden mit ihren Ausgangswerten in das Modell eingegeben. In der Folge unterliegen sie verschiedenen Arten von Veränderungen (Abb. 3): Einwohner bekommen Kinder, ziehen um, zu oder fort, altern und sterben, die Anzahl der Beschäftigten nimmt zu oder ab, ihre Verteilung auf die einzelnen Wirtschaftsbereiche ändert sich mit dem Strukturwandel der Wirtschaft, Wohnungen werden neu- oder umgebaut, abgerissen oder in Büros umgewandelt, Flächen werden anderen Nutzungen zugeführt, Straßen oder öffentliche Verkehrslinien werden neu- oder ausgebaut, auf anderen Linien wird der Verkehr verstärkt, verringert oder eingestellt.

Abfolge und Ausmaß der Veränderungen werden durch Vorgaben gesteuert, die für jeden Simulationslauf frei gewählt werden können. Die Vorgaben enthalten Annahmen über die Entwicklung technischer Planungsgrößen und Richtwerte, Kosten- und Finanzierungsdaten sowie Parameter, die, Ergebnis modellexterner empirischer Untersuchungen, Lage und Verlauf einzelner Modellfunktionen festlegen. Zu den Vorgaben gehören auch die Ergeb-

nisse des ersten Teilmodells, des (Makro-)Prognosemodells, bzw. bis zu dessen Fertigstellung Annahmen über den Gesamteinwohner- und Beschäftigtenzuwachs des Untersuchungsgebiets sowie den Einwohner- und Beschäftigtenzuwachs der Umlandzonen. Die wichtigsten Vorgaben jedoch sind die Planungsmaßnahmen.

Planungsmaßnahmen sind alle bewußten Eingriffe der Stadt oder eines anderen Planungsträgers in die Stadtentwicklung. Gegenwärtig unterscheidet das Modell drei Gruppen von Planungsmaßnahmen. Die erste Gruppe bilden institutionelle Maßnahmen, d.h. in erster Linie Maßnahmen der Bauleitplanung wie Festlegung von Art und Maß der zulässigen Nutzung in jeder Zone. Diese Flächenausweisungen bedeuten allein noch keine eigentliche Investition oder Bautätigkeit, sie setzen lediglich den Rahmen für die im Modell simulierten Entscheidungen der privaten Investoren und Bauherren. Die zweite Gruppe von Planungsmaßnahmen sind direkte Eingriffe der Stadt oder eines anderen Planungsträgers in die Flächennutzung der Zonen. Das Modell verarbeitet ca. 60 verschiedene Arten von Planungsmaßnahmen auf den Gebieten

- Wohnungsbau,
- Gewerbeansiedlung,
- Gemeinbedarfseinrichtungen,
- Erholungseinrichtungen,
- Einkaufseinrichtungen,
- Ruhender Verkehr,
- Naherschließung,
- Ver- und Entsorgung,
- Städtische Baulandreserve,
- Stadterneuerung.

Bei allen Maßnahmen kann für jede Zone ein eigenes, zeitlich abgestuftes Bau- und Durchführungsprogramm vorgegeben werden. Die dritte Gruppe von Planungsmaßnahmen schließlich sind die Verkehrsmaßnahmen. Das Modell erlaubt die Eingabe beliebiger, zeitlich abgestufter Programme zum Ausbau der öffentlichen Verkehrsmittel und des Straßennetzes, gegebenenfalls auch

die experimentelle Einführung neuartiger, bisher nicht vorhandener Verkehrssysteme (Abb. 4, 5).

Die Simulation einer Periode beginnt beim Verkehrsangebot am Anfang der Periode. Aus den Ausgangsnetzen und den Verkehrsbauten der zurückliegenden Perioden ergeben sich die Verkehrsnetze der Periode. Aufgrund des Verkehrsangebots und der vorhandenen Flächennutzungsverteilung werden Zugänglichkeitswerte aller Teilzonen berechnet. Aus Zugänglichkeit und anderen Attraktivitätsmerkmalen der Zonen wird die wahrscheinliche Verteilung der für die betreffende Periode erwarteten Bautätigkeit auf die zur Verfügung stehende Baulandreserve ermittelt. Es folgt die Fortschreibung des Einwohner- und Beschäftigtenbestandes der Zonen unter Berücksichtigung der Bautätigkeit, der natürlichen Bevölkerungsentwicklung, der Veränderung der Flächenanforderungen und Belegungsziffern sowie die Berechnung sämtlicher Folgeflächen. Dann kann aus den mittleren Einwohner- und Beschäftigtenzahlen der Periode das Verkehrsaufkommen sowie die wahrscheinliche Verteilung der Verkehrsströme auf die Teilstrecken beider Verkehrsarten berechnet werden. Damit ist die Simulation einer Periode beendet. Alle Inventardaten haben neue Zustandswerte erhalten, die nächste Periode kann beginnen.

2.4 Programmaufbau

Das Programm des Polis-Modells besteht aus zwei selbständigen Hauptprogrammen mit je mehreren Unterprogrammen (Abb. 6). Das erste Programm POLIS 1 enthält die Programnteile zur Analyse einer Verkehrsvariante. Es erzeugt ein Übergabeband mit allen notwendigen Daten sämtlicher Ausbaustufen beider Netze einer Variante. Mit diesem Band als Eingabeband wird im zweiten Programm POLIS 2 die eigentliche Simulation der Stadtentwicklung durchgeführt.

Das Netzanalyseprogramm folgt im wesentlichen gebräuchlichen Verfahren der Verkehrsplanung. Zunächst wird aus den Strecken-

daten des Grundnetzes und etwaiger Netzveränderungen das zu untersuchende Netz zusammengestellt. Dieses Netz wird in Form der sogenannten Netzbeschreibung für die Weiterverarbeitung aufbereitet. Die Netzbeschreibung ist die Grundlage für die Ermittlung der "Bäume" der kürzesten Wege aller internen zu allen internen und externen Zonen sowie der für die Wege benötigten Reisezeiten. Das Verfahren wird für beide Netze in gleicher Weise angewendet. Ein Unterschied besteht lediglich darin, daß im Netz der öffentlichen Verkehrsmittel bei der Wahl des kürzesten Weges die je nach Linienführung und Zugfolge unterschiedlichen Umsteigezeiten berücksichtigt werden müssen. Bei den Pkw-Fahrten dagegen werden die erforderlichen Zeiten für das Aufsuchen eines Parkplatzes in der Zielzone zu den eigentlichen Reisezeiten hinzugerechnet.

Die Reisezeiten sind die Grundlage für die Berechnung der Zugänglichkeitswerte der Zonen. Reisezeitfaktoren als Funktion der Reisezeit drücken den Widerstand aus, der auf dem jeweiligen Verkehrsnetz zwischen zwei Zonen zu überwinden ist. Je kleiner dieser Widerstand ist, desto mehr trägt eine Zone zur Zugänglichkeit der anderen Zone bei. Andererseits hängt der Beitrag einer Zone zur Zugänglichkeit einer anderen von der Größe der Aktivitäten in ihr ab.

Als nächstes erfolgt die Simulation der Flächennutzungsentwicklung. Der Einwohnerzuwachs einer Zone setzt sich zusammen aus natürlicher Bevölkerungsentwicklung und Wanderungssaldo. Mit Hilfe der für die Gesamtstadt prognostizierten Wohnraumbelegung läßt sich der Gesamtzuwachs an Wohnraum in einer Periode berechnen, der erforderlich ist, wenn der vorgegebene Gesamteinwohnerzuwachs realisiert werden soll. Die Verteilung des Gesamtwohnraumzuwachses auf die Zonen erfolgt, wie alle Verteilungsoperationen des Modells, proportional zu dem jeweiligen "Entwicklungspotential" der Zone. Das Entwicklungspotential ist allgemein die mit der Attraktivität der Zone gewichtete Kapazität, im Falle des Wohnraumzuwachses die Flächenreserve der Zone, gewichtet mit einer Funktion der Zu-

gänglichkeit und anderer Attraktivitätsmerkmale. Ist der Zuwachs an Wohnräumen bekannt, kann die Einwohnerkapazität der Zone am Ende der Periode berechnet werden. Dabei ist nicht nur das Absinken der Wohnraumbelegung im Stadtdurchschnitt zu berücksichtigen, sondern auch die für ein Wohngebiet mit einer bestimmten Altersstruktur und einem bestimmten Baualter und Zustand der Wohnungen charakteristischen Abweichung vom Stadtdurchschnitt.

Auf ähnliche Weise wird der Zuwachs an Arbeitsplätzen im produzierenden Gewerbe und im Dienstleistungsgewerbe auf die Zonen verteilt. Anschließend werden mit Hilfe spezifischer Flächenbeiwerte die Bruttogeschoßflächen neu zu errichtender Wohn- und Nichtwohngebäude sowie das erforderliche Nettobauland, die Flächen für Gemeinbedarfseinrichtungen, die Grünflächen und Erschließungsflächen berechnet. Für den Fall, daß die Nachfrage nach Bauland das Angebot in einer Zone übersteigt, können besondere Maßnahmen zum Ausgleich der Übernachtungsvorgesehen werden. Es folgen die Fortschreibung der Einwohneraltersgruppen, der Beschäftigten in den einzelnen Wirtschaftsbereichen, der Wohnungen und Flächen sowie die Aufstockung der Flächenreserve durch Abriß vorhandener Bausubstanz.

Dann kann das Verkehrsaufkommen der Periode berechnet werden. Die Fahrten in beiden Richtungen zwischen zwei Zonen werden dabei direkt aus den mittleren Einwohner- und Beschäftigtenzahlen der Periode unter Benutzung der Reisezeitfaktoren berechnet. Der Anteil der Fahrten mit öffentlichen Verkehrsmitteln zwischen zwei Zonen wird nach dem Reisezeitverhältnis geschätzt, wobei die Modal-Split-Funktion von Periode zu Periode entsprechend dem Ansteigen des Motorisierungsgrads verschoben wird.

Die Fahrten werden abschließend auf die Teilstrecken der ermittelten kürzesten Wege umgelegt. Durch Zusammenführen von Streckendaten und Teilstreckenbelastungen können Leistungsdaten der beiden Verkehrsnetze wie Ausnutzung des Platzangebots bzw.

Kapazitätsauslastung der Teilstrecken, Personenkilometer und Reisezeiten nach Strecken- bzw. Straßenarten ermittelt werden. Kapazitätsüberschreitungen im Straßenverkehr haben Stauungen zur Folge, denen durch Zuschlag zu den Reisezeiten Rechnung getragen wird.

Die Ausgabe der Simulationsergebnisse jeder Periode erfolgt auf dem Zeilendrucker in Form von Tabellen und Diagrammen. Außerdem kann in jedem Stadium der Simulation die räumliche Verteilung beliebiger Modellvariablen in Karten ausgegeben werden (Abb. 7). Nach Durchlaufen aller Perioden wird die Entwicklung der Flächennutzung und des Verkehrsaufkommens der Gesamtstadt sowie einer Reihe von Kennziffern zur Bewertung der Flächennutzungs- und Verkehrsstruktur der Stadt in Tabellen und Diagrammen dargestellt.

Daneben werden die wichtigsten durch die untersuchte Variante verursachten einmaligen und wiederkehrenden Einnahmen und Ausgaben der Stadt und anderer beteiligter Gruppen (Bund/Land/Kreis, Verkehrsbetriebe, Verkehrsbenutzer, Bauherren, Mieter) errechnet und in einer Matrix der Zahlungsströme der einzelnen Perioden dargestellt. Abschließend werden die Zahlungsstromsalden der einzelnen Gruppen, auf das Ausgangsjahr bezogen, abgezinst und akkumuliert und mit den entsprechenden Salden einer hypothetischen "Nullvariante" verglichen. Die Nullvariante ist definiert als "Entwicklung ohne Planungseingriffe" und dient als Vergleichsbasis für alle Planungsvarianten.

3. Bewertung der Ergebnisse

Es ist die Aufgabe des POLIS-Modells, Basisinformationen für die vergleichende Bewertung der Planungsvarianten zu liefern. Implizite Wertungsvorgänge im Planungsprozeß sollen durch logisch konsistente und nachvollziehbare Verfahren überlagert werden.

Die Beziehungen zwischen den untersuchten Maßnahmenkombinationen und ihren Konsequenzen wird durch das Simulationsmodell

hergestellt. Die quantitative Erfassung der Auswirkungen einer Variante im Simulationsmodell soll ein differenziertes Bild nicht nur der technischen, sondern auch der ökonomischen und sozialen Aspekte dieser Auswirkungen ergeben. In der Nutzenbilanz zum Beispiel sind die Auswirkungen des Verkehrssystems der untersuchten Variante für verschiedene Gruppen dargestellt (Abb. 8): die Benutzer des Verkehrssystems, die Einwohner insgesamt sowie Minderheiten wie alte und arme Einwohner, die Betriebe und Institutionen und schließlich die Umwelt. Die Auswirkungen werden durch Kennziffern ausgedrückt, deren Entwicklung über die Simulationsperioden verfolgt werden kann. So wird die Qualität des Verkehrssystems aus der Sicht der Benutzer durch Merkmale wie Geschwindigkeit, Stauzeit, Wartezeit, Fußweg und Sitzplatzangebot bestimmt. Für die Einwohner, Betriebe und Institutionen äußert sie sich in der Zugänglichkeit zu ausgewählten Zielaktivitäten. Zugänglichkeitsmaße werden außerdem gesondert für die Wohnstandorte der alten und armen Einwohner unter ausschließlicher Berücksichtigung der Versorgung mit öffentlichen Nahverkehrsmitteln berechnet. Die externen Auswirkungen des Verkehrssystems auf die Umwelt und die damit verbundenen sozialen Kosten sind durch Meßgrößen wie Luftverschmutzung, Lärmbelästigung, Behinderung, Unfälle und Sachschäden erfaßt.

Die finanziellen Implikationen der Planungsvarianten dienen als eine weitere Bewertungsgrundlage. In ihnen kommen die Realisierbarkeit der Variante sowie die Relation zwischen Aufwand und Ergebnis der Planung zum Ausdruck. Durch die differenzierte Angabe der Verteilung der Gesamtkosten auf die einzelnen Gruppen der Stadt wird die Frage, welchen Interessen die Planungsvariante nützt, in das Bewußtsein gerückt.

Zweck des eigentlichen Bewertungsmodells wird es zunächst sein, die vielfältigen, verschieden dimensionierten und skalierten Auswirkungen der Planungsvarianten vergleichbar darzustellen. Dem Entscheidungsträger sollen die unverfälschten Implikationen der verschiedenen Varianten, geordnet und ver-

gleichbar gegenübergestellt, zur Interpretation überlassen werden. Um jedoch aus der Fülle der dabei anfallenden Daten eine eindeutige und nachvollziehbare Rangfolge der Varianten hinsichtlich ihres Nutzens ableiten zu können, sollen darüberhinaus die verschiedenen Nutzeneffekte einer Variante in einem Gesamtnutzenmaß zusammengefaßt werden. Dies bedingt eine Gewichtung der Planungsziele sowie die Zuordnung von Nutzenfunktionen zu den einzelnen Merkmalen der Varianten. Die Bewertung soll in den verschiedenen Stadien der Planung wiederholt vorgenommen werden.

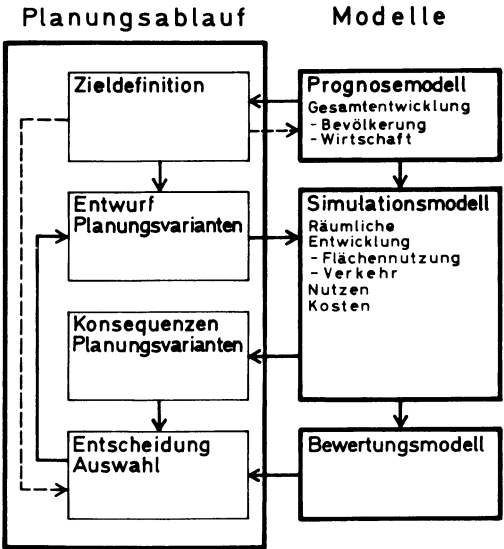


Abb. 1 Planungsablauf und Modelle

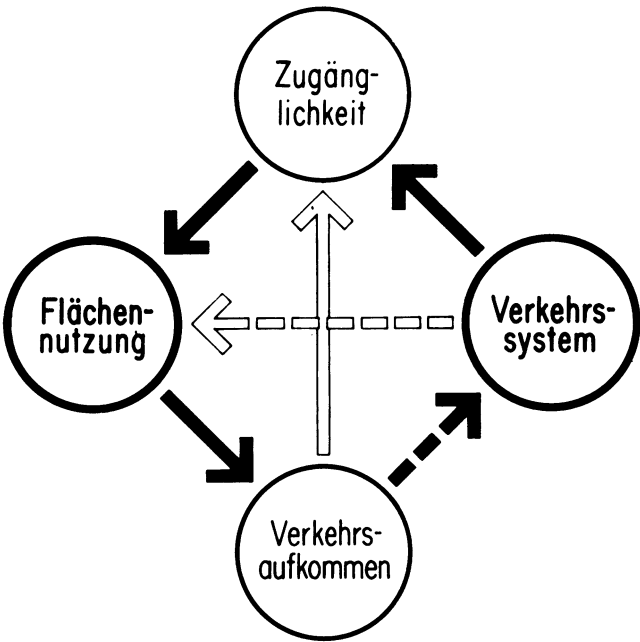


Abb. 2 Flächennutzung und Verkehrssystem

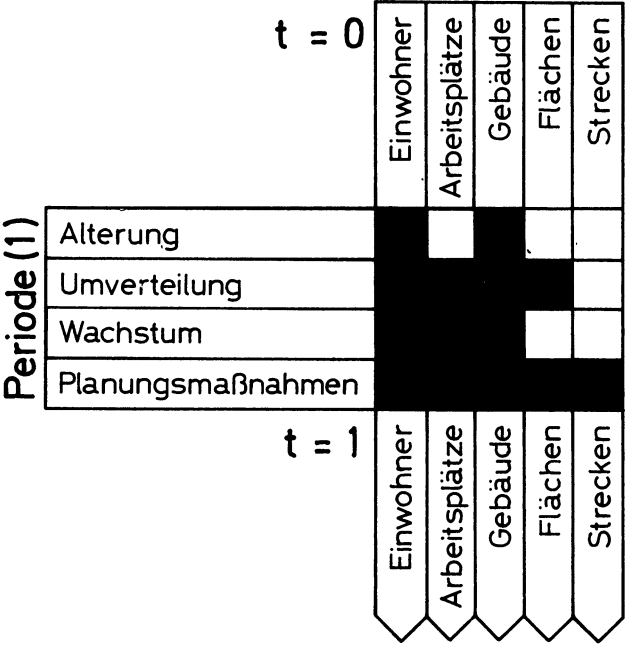


Abb. 3 Inventardaten und ihre Veränderung im Ablauf einer Simulationsperiode

1 WOHNUNGSRAU						2 GEWERBEANSIEDLUNG						3 GEMEINBEDARFSEINRICHTUNGEN I								
Z	M	(1)	(2)	(3)	(4)	(5)	Z	M	(1)	(2)	(3)	(4)	(5)	Z	M	(1)	(2)	(3)	(4)	(5)

4 GEMEINBEDARFSEINRICHTUNGEN II						5 ERHOLUNGSEINRICHTUNGEN						6 EINKAUFSEINRICHTUNGEN								
Z	M	(1)	(2)	(3)	(4)	(5)	Z	M	(1)	(2)	(3)	(4)	(5)	Z	M	(1)	(2)	(3)	(4)	(5)

7 RUHENDER VERKEHR						8 HAARSCHLIESSUNG						9 VER- UND ENTSORGUNG								
Z	M	(1)	(2)	(3)	(4)	(5)	Z	M	(1)	(2)	(3)	(4)	(5)	Z	M	(1)	(2)	(3)	(4)	(5)

10 STÄDTISCHE BAULANDRESERVE						11 STÄDTISIERUNG						12								
Z	M	(1)	(2)	(3)	(4)	(5)	Z	M	(1)	(2)	(3)	(4)	(5)	Z	M	(1)	(2)	(3)	(4)	(5)

Z = Zone
 M = Maßnahmentyp

Abb. 4 Planungsmaßnahmen zur Flächennutzung: Eingabeformular (Schema)

UNRAUSTRECKEN (OPNV)

P	A	B	E	T/G	L	FQ	Z	BJ

UNBAUSTRECKEN (KW)

P	A	B	E	T/G	Z	BJ

NEUBAUSTRECKEN (OPNV)

P	A	B	E	T/G	L	FQ	Z	BJ

NEUBAUSTRECKEN (PKW)

P	A	B	E	T/G	Z	BJ

AUFGELASSENE STRECKEN (OPNV)

P	A	B	E	T/G	L	FQ	Z	BJ

AUFGELASSENE STRECKEN (PKW)

P	A	B	E	T/G	Z	BJ

P = Periode

A = A-Knoten

B = B-Knoten

E = Länge

T/G = Zeit/Geschwindigkeit

L = Linien

FQ = Zugfolge

Z = Zone

BJ = Baujahr

Abb. 5 Verkehrsbaumaßnahmen: Eingabeformular (Schema)

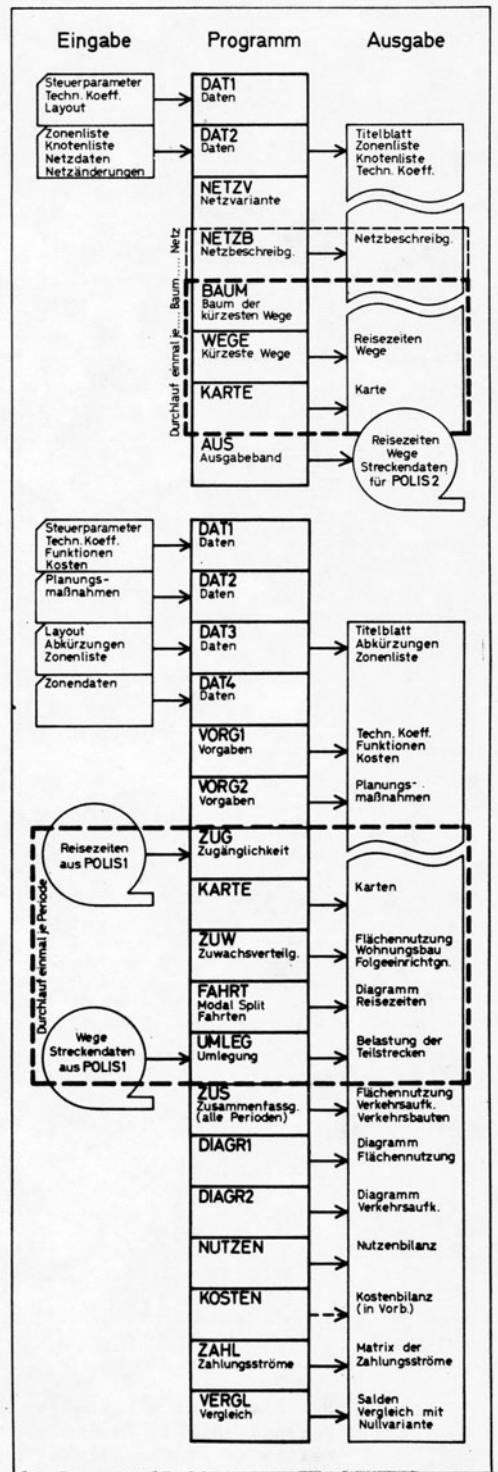


Abb. 6 Programmaufbau des POLIS-Modells:
POLIS 1 (oben)
POLIS 2 (unten)

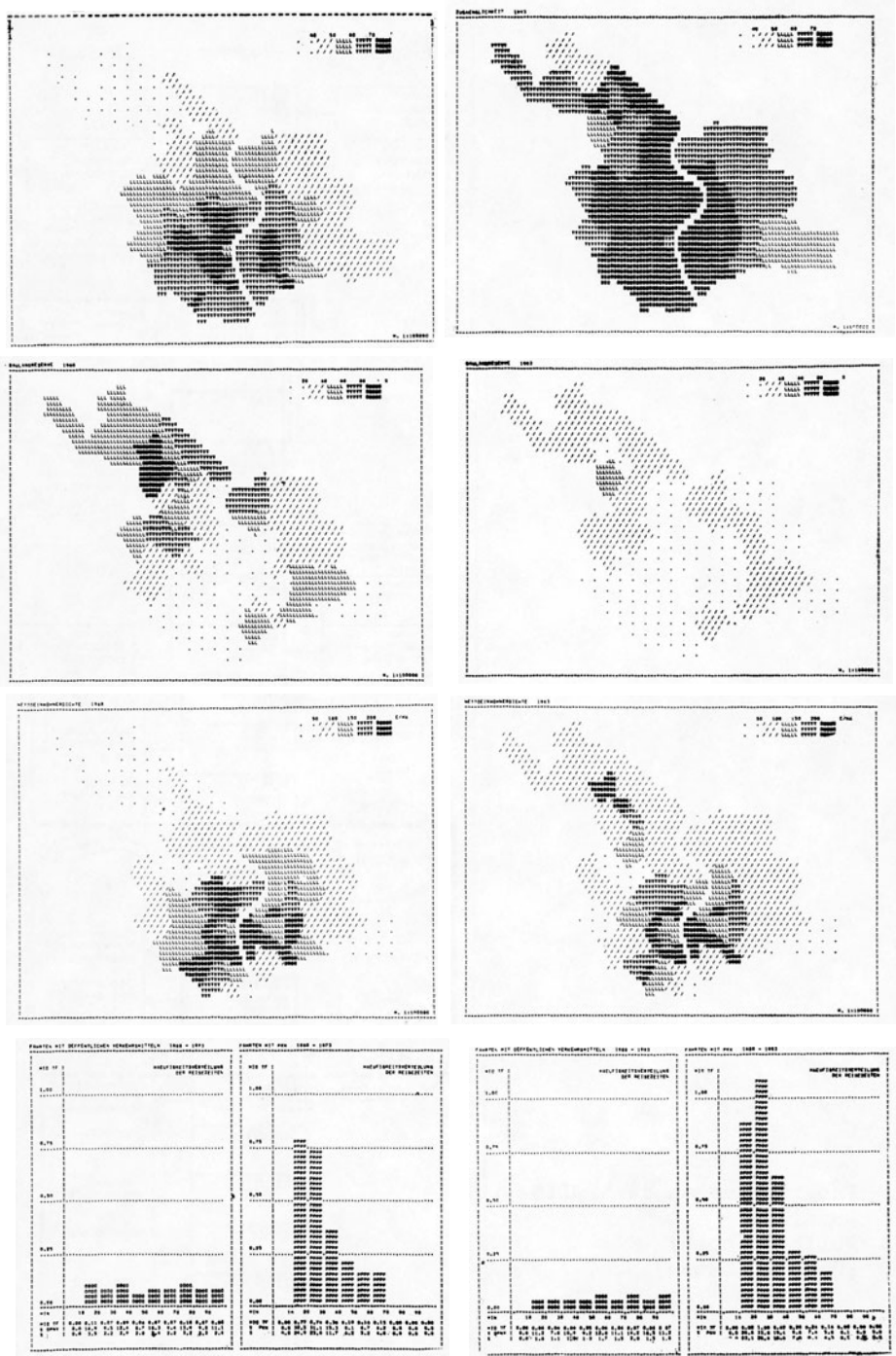


Abb. 7 Beispiele für Ausgabe der Simulationsergebnisse (Köln):
Zugänglichkeit, Baulandreserve, Einwohnerdichte, Reise-
zeiten zu Beginn (links) und am Ende (rechts) der Simu-
lation

**Ein Modell zur optimalen Wärmeversorgung in Städten
durch leitungsgebundene Energieträger**
von K. P. Liesenfeld, Aachen

1. Einführung

Im Rahmen der allgemeinen Bemühungen, die Verunreinigungen unserer Umwelt zu verringern, sind Maßnahmen zur Reinhaltung der Luft von besonderer Bedeutung. Im Raumwärmesektor bietet sich dazu (bei genügend hoher Wärmedichte des betrachteten Gebietes) die Verwendung leitungsgebundener Energieträger an, da diese einen weitgehend immissions- und emissionsfreien Betrieb von Wärmeanlagen ermöglichen.

Man nennt Energieträger leitungsgebunden, wenn sie zum Transport an den Endverbraucher fest installierte Systeme - Netze - benötigen: also Gas, Heizwasser, Dampf und elektrischer Strom.

Ein Energieversorgungsunternehmen (EVU), dem die Versorgung eines räumlich abgegrenzten Gebietes - etwa einer Stadt - mit den Energieträgern Gas, Heizwasser und elektrischer Strom obliegen möge, kann nun vor der folgenden Problemstellung stehen:

Das EVU will innerhalb seines Versorgungsgebietes in den derzeit nicht mit leitungsgebundenen Energieträgern versorgten Wärmemarkt expandieren, ohne jedoch einen Teil seiner bisherigen Abgabe an leitungsgebundenen Energieträgern, der mit Grundlast bezeichnet sei, zu beeinträchtigen und ohne sein augenblicklich bestehendes Energieverteilungs- und Beschaffungssystem ausbauen zu müssen. Die einzigen Investitionen, die es bezüglich seiner Expansion tätigen will, seien die, welche zur Erstellung von Hausanschlüssen notwendig sind.

Das Modell ist Bestandteil eines Forschungsauftrages, der am Lehrstuhl für Unternehmensforschung an der RWTH Aachen mit Unterstützung der WIBERA A.G. durchgeführt wurde.

Der augenblicklich nicht mit leitungsgebundenen Energieträgern bediente Teil des Wärmemarktes sei mit Wärmepotential bezeichnet. Dieses Wärmepotential ist örtlich qualitativ und quantitativ bestimmt durch:

- a) eine mögliche Substitution von nichtleitungsgebundenen Energieträgern, die einen Teil des bereits bestehenden Wärmebedarfs decken, durch leitungsgebundene,
- b) eine mögliche Bedienung des noch latenten Wärmebedarfs mit leitungsgebundenen Energieträgern.

Das so definierte Wärmepotential soll sich in der Hauptsache auf Raumheizungszwecke beschränken.

Die vom EVU angestrebte Absatzexpansion soll so erfolgen, daß sie unter Einhaltung gewisser Nebenbedingungen, die später aufgeführt werden, einen maximalen Periodengewinn bei konstanten Preisforderungen erbringt.

Zur Bedienung des Wärmepotentials mit leitungsgebundenen Energieträgern sollen Gas, Heizwasser und elektrischer Strom zum Schwachlasttarif (Schwachlaststrom) zur Verfügung stehen.

2. Skizze des mathematischen Modells

2.1 Darstellung der räumlichen Verteilung des Wärmepotentials im Modell

Zur Darstellung der räumlichen Verteilung des Wärmepotentials wird das Versorgungsgebiet in $i (i=1, \dots, n)$ Zellen unterteilt. Diesen Zellen werden dann die entsprechenden Einzelabnehmer und deren Wärmepotentiale zugeordnet. Damit bestimmt sich das Wärmepotential a_{ib} ($i=1, \dots, n; b=1, \dots, 4$) einer Zelle i aus der Aggregation der Wärmepotentiale aller einer Zelle zugeordneten Einzelabnehmer, wobei der Index $b (b=1, \dots, 4)$ zwischen den Abnehmergruppen, in die die Einzelabnehmer eingeteilt werden können, differenziert. Der Index b kennzeichnet mit:

b=1: Abnehmer, die Wärme zu Zwecken der Einzelheizung
b=2: Abnehmer, die Wärme zu Zwecken der Etagencentralheizung
b=3: Abnehmer, die Wärme zu Zwecken der Kellercentralheizung
b=4: Abnehmer, die Wärme zu sonstigen Zwecken
 benötigen.

Aus Abb. 1 können die Beschränkungen und die Bedienungsmöglichkeiten des Wärmepotentials einer Zelle, nach Abnehmergruppen getrennt, für die einzelnen Energieträger l ($l=1,2,3$) entnommen werden, wenn man dem Schema in Abb. 1 folgende Interpretation zugrundelegt:

"1" : Bedarfsdeckung mit Energieträger l bei der Verbrauchergruppe b möglich

"0" : Bedarfsdeckung nicht möglich.

Für Wärmebedarfsdeckung möglicher leitungsge- bundener Energieträger		Abnehmergruppe			
		b=1	b=2	b=3	b=4
Gas	$l = 1$	1	1	1	1
Heizwasser	$l = 2$	0	0	1	1
Schwachlast- strom	$l = 3$	1	0	0	1
Wärmepotentiale a_{ib} in der Zelle i		a_{i1}	a_{i2}	a_{i3}	a_{i4}

Abb. 1
 Beschränkungen und Bedienungsmöglichkeiten des Wärmepotentials
 einer Zelle

Mit Hilfe von Abb. 1 sei nun noch die Matrix $R = (r_b^1)_{3,4}$ definiert, mit

$$r_b^1 = \begin{cases} 1, & \text{falls Bedarfsdeckung mit Energieträger 1 bei} \\ & \text{der Verbrauchergruppe b möglich} \\ 0, & \text{sonst.} \end{cases} \quad (1)$$

2.2 Die Zielfunktion des Modells

Um die Zielfunktion darstellen zu können, werden zunächst die Variablen

$$y_{ib}^1, \quad y_{ib}^1 \geq 0 \quad \begin{matrix} (i=1, \dots, n) \\ (l=1, 2, 3) \\ (b=1, \dots, 4) \end{matrix} \quad (2)$$

eingeführt. Sie beschreiben die Wärmeleistung, die der Abnehmergruppe b zur Bedienung des in der Zelle i bestehenden Wärmepotentials a_{ib} mit der Lieferung des Energieträgers l zur Verfügung gestellt wird. Führt man jetzt noch die Preiskoeffizienten ¹⁾

$$c_{ib}^1 \quad \begin{matrix} (i=1, \dots, n) \\ (l=1, 2, 3) \\ (b=1, \dots, 4) \end{matrix} \quad (3)$$

ein, so läßt sich die zu maximierende Zielfunktion schreiben zu:

$$Z(y_{ib}^1) = \sum_{i=1}^n \sum_{b=1}^4 \sum_{l=1}^3 c_{ib}^1 r_b^1 y_{ib}^{1l}. \quad (4)$$

Die Preiskoeffizienten c_{ib}^1 setzen sich aus folgenden Bestandteilen zusammen:

¹⁾ Bei Gas muß wegen unterschiedlich hoher Erlöse eigentlich zwischen den Kundenarten "bisheriger Gaskunde" und "bisheriger Nichtgaskunde" unterschieden werden. Darauf wird hier jedoch aus Gründen einer vereinfachten Darstellung verzichtet.

- a) arbeits- und leistungsabhängigen Erlösen,
- b) arbeitsabhängigen Beschaffungskosten,
- c) erfolgswirksamen Auflösungen von Baukostenzuschüssen,
die der Kunde bei Errichtung eines Hausanschlusses evtl.
an das EVU zu zahlen hat,
- d) jährlichen Abschreibungen des EVU's auf die für einen
Hausanschluß erstellten Anlagen,
- e) Transportkosten der einzelnen Energieträger.

Auf eine nähere Beschreibung der Art und Weise wie diese Bestandteile in den Preiskoeffizienten berücksichtigt werden, soll hier verzichtet werden.

2.3 Die Restriktionen des Modells

2.3.1 Bedienung der den Zellen zugeordneten Wärmepotentiale

Die einzelnen Energieträger l ($l=1,2,3$) können einander bei der Bedienung des in der Zelle i ($i=1,\dots,n$) bestehenden Wärmepotentials a_{ib} ($b=1,\dots,4$) innerhalb des in Abb. 1 dargestellten Rahmens substituieren. Nimmt man für die bei den einzelnen Abnehmergruppen b eingesetzten Energieumwandlungssysteme einen Wirkungsgrad η_b^1 an, so läßt sich die folgende Nebenbedingung formulieren:

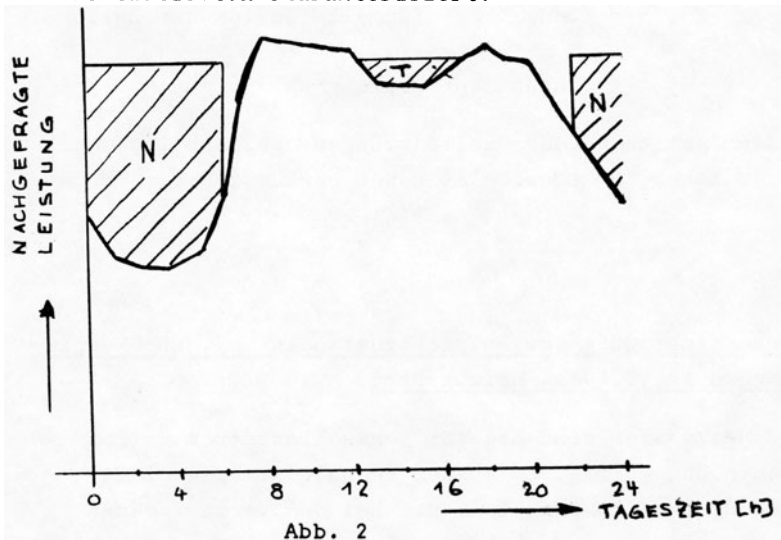
$$\sum_{l=1}^3 \eta_b^1 r_b^l y_{ib}^l \leq a_{ib} \quad \begin{matrix} (i=1,\dots,n) \\ (b=1,\dots,4) \end{matrix} \quad (5)$$

2.3.2 Berücksichtigung der Betriebszustände der Netze

Die zum Transport der leitungsgebundenen Energieträger benutzten Netze sind im allgemeinen vermascht und kapazitätsbeschränkt. Da die Abnehmer gewisse Mindestpotentiale (Drücke bzw. Spannungen) bezüglich der gelieferten Energieträger fordern und die Potentiale vom jeweiligen Betriebszustand des Netzes abhängig sind, müssen diese Betriebszustände im Modell berücksichtigt werden. Ein Betriebszustand heie im folgenden zulssig, wenn dieser Betriebszustand technisch mglich ist und auerdem die Mindestpotentialbedingungen erfllt.

2.3.2.1 Berücksichtigung des Betriebszustandes beim Energietrger 1=3 (Schwachlaststrom)

Eine nhere Erluterung des Begriffs Schwachlaststrom ist mit Hilfe von Abb. 2 mglich, die eine qualitativ typische Belastungskurve darstellt. Diese Belastungskurve ist neben ihrer Spitzenlast durch ihre Belastungstler whrend sogenannter Schwachlastzeiten charakterisiert.



Belastungskurve bei elektrischem Strom

Mit Schwachlaststrom (oder Nacht- und Tagesnachladestrom bei Raumheizung) wird nun die elektrische Leistung bezeichnet, die man zur Auffüllung der Belastungstäler (zu in der Regel verhältnismäßig niedrigem Preis), zusätzlich verkauft. Die in Abb. 2 schraffierte Fläche beschreibt dann die Schwachlastarbeit.

Setzt man voraus, daß der Betriebszustand der elektrischen Netze auch während der Verteilung der jährlichen Spitzenlast zulässig ist, dann kann man zur Sicherstellung eines zulässigen Betriebszustandes bei Schwachlaststrom folgendermaßen vorgehen:

Man faßt die Zellen $i(i=1,\dots,n)$ eines Versorgungsgebietes so zu Zellengruppen $j(j=1,\dots,J^3)$ zusammen, daß man jeder dieser Zellengruppen eine sekundäre Quelle mit der Quelleistung Q_j^3 zuordnen kann. Zur Sicherung eines zulässigen Betriebszustandes genügt dann die folgende Nebenbedingung:

$$\sum_{i \in I_j^3} \sum_{b=1,4} y_{ib}^3 \leq Q_j^3 - z_j^3, \quad (j=1,\dots,J^3), \quad (6)$$

mit z_j^3 = Grundlast innerhalb der Zellengruppe j ,
 I_j^3 = Indexmenge für alle Zellen der Zellengruppe j ,
 J^3 = Anzahl der Zellengruppen,

wobei bei der Festlegung der Quelleistungen die höchstzulässige Erwärmung der Kabel für Dauerbelastungen berücksichtigt werden muß.

2.3.2.2 Berücksichtigung des Betriebszustandes bei den Energieträgern 1=1,2 (Gas, Heizwasser)

Bei Gas und Heizwasser sind die für Schwachlaststrom getroffenen Annahmen nicht gültig. Es müssen deshalb, um die Mindestpotentialbedingungen zu erfüllen, die bei den verschiedenen

Belastungsfällen an den Knotenpunkten der Netze auftretenden Drücke explizit im Modell berechnet werden. Zu diesem Zweck müssen Netzberechnungen in das Modell integriert werden.

Zur Berechnung eines Netzes mit N Kanten und K Knotenpunkten, wovon E Einspeisestellen seien, kann man zur Bestimmung der

$$N + E + K - E = N + K \quad (7)$$

unbekannten Flüsse und Drücke:

- a) K Gleichungen aus den Kirchhoff'schen Knotenpunktsbedingungen mit $N + E$ unbekannten Flüssen,
- b) N Gleichungen aus den Druckverlusten der N Einzelrohre des Netzes mit $K - E$ unbekannten Drücken,
- c) die Belastung aller Knotenpunkte,
- d) die Rohrnetzdaten,
- e) die Einspeisedrücke

heranziehen.

Zur Aufstellung der einer integrierten Netzberechnung entsprechenden Nebenbedingungen wird nun zunächst jedem Knotenpunkt eines Netzes eine Zelle $i(i=1, \dots, n)$ zugeordnet.

Netze zum Transport von Gas und Heizwasser lassen sich in die Klasse der ungerichteten, endlichen und einfachen Graphen einordnen. Zur Berechnung der Betriebszustände müssen auf den Kanten der Netze jedoch Bezugsrichtungen zur Orientierung der auftretenden Flüsse gewählt werden. Diesem System von Bezugsrichtungen werde nun ein gerichteter Graph mit der Inzidenzmatrix

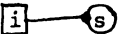
$$T^1 = (t_{sj}^1)_{K^1, N^1} \quad (l=1,2) \quad (8)$$

zugeordnet, wobei

- j = Index für alle Kanten,
- s = Index für alle Knotenpunkte,
- K^1 = Anzahl der Knotenpunkte im Netz des Energieträgers 1,
- N^1 = Anzahl der Kanten im Netz des Energieträgers 1.

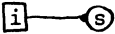
Damit führt die Formulierung der Kirchhoff'schen Knotenpunktbedingungen auf die Nebenbedingungen:

$$\sum_{b=1}^4 y_{ib}^1 r_b^1 - F^1 \sum_{j=1}^{N^1} t_{sj}^1 q_j^1 = -z_i^1, \quad \begin{matrix} (l=1,2), \\ (s=1,\dots,K^1), \end{matrix} \quad (9)$$

mit ,

wobei F^1 = konstanter Faktor zur Umrechnung der Flüsse in Wärmeeinheiten

q_j^1 = Variable zur Darstellung des in der Kante j auftretenden Flusses beim Energieträger 1

ist, und die Zuordnung  definiert sei durch:

$$\text{Diagram} : : = \text{Zelle } i \text{ ist Knotenpunkt } s \text{ zugeordnet.} \quad (10)$$

Bei E^1 ($l=1,2$) Einspeisestellen lassen sich die folgenden Nebenbedingungen formulieren:

$$p_s^1 = P_{Es}^1 \quad (s=K^1-E^1-1, K^1-E^1-2, \dots, K^1) \quad (11)$$

$(l=1,2),$

mit $p_s^1 \geq 0$, wobei

P_{Es}^1 = Konstante zur Darstellung des Einspeisedruckes am Knotenpunkt s für Energieträger 1,

p_s^1 = Variable zur Darstellung des Betriebsdruckes am Knotenpunkt s für Energieträger 1.

Nimmt man nun noch die Druckverlustgleichungen für alle in den Netzen auftretenden Einzelrohre in das Modell auf, dann ist der Betriebszustand der Netze vollständig beschrieben, Ohne jetzt eine Aussage über die Form dieser Druckverlustgleichungen machen zu wollen, können dieselben formal folgendermaßen beschrieben werden durch:

$$p_s^1 - p_{s+1}^1 = f_j^1(q_j^1) \quad , \quad (12)$$

wenn die Kante j eines Netzes mit den Knotenpunkten s und $s+1$ inzident ist. Mit (12) und (8) kann jetzt die folgende Nebenbedingung zur Bestimmung der Druckverluste im Einzelrohr formuliert werden:

$$p_s^1 - p_{s+1}^1 - t_{sj}^1 f_j^1(q_j^1) = 0 \quad (l=1,2) \quad , \quad (13)$$

$$(j=1, \dots, N^1) \quad .$$

Um jetzt schließlich einen zulässigen Betriebszustand sicherzustellen, genügt es, für entsprechende Knotenpunkte die Einhaltung gegebener Mindestpotentiale P_{\min} zu fordern, und die gesamte dem System zur Bedienung der Wärmepotentiale verfügbare Leistung zu beschränken.

Also müssen noch die Ungleichungen

$$p_s^1 \geq P_{\min s}^1 \quad , \quad (s=1, \dots, K^1 - E^1) \quad (14)$$

$$(l=1,2)$$

und

$$\sum_{i=1}^n \sum_{b=1}^4 r_b^1 y_{ib}^1 = \sum_{u^1} Q_{u^1}^1 - \sum_{i=1}^n z_i^1 \quad (l=1,2) \quad (15)$$

$$\text{mit } u^1 \in U^1 \quad \text{und} \quad |U^1| = E^1 \quad ,$$

wobei U^1 = Indexmenge aller Einspeisestellen
des Energieträgers 1
 Q_{u^1} = Konstante zur Darstellung der maxi-
malen Kapazität der Einspeisestelle
 u^1

in das Modell ist, aufgenommen werden. Damit ist das gesamte Modell formal beschrieben.

2.4 Struktur und Klassifikation des Modells

Die Struktur des Modells ist blockartig, da die einzelnen Energieträger in voneinander unabhängigen Netzen transportiert werden.

Weiterhin ist das Modell linear in allen Variablen, mit Ausnahme derer, die zur Berechnung der Druckverluste in Einzelrohren eingeführt wurden. Da diese in Gleichheitsnebenbedingungen auftreten, kann das Modell als i.a. nichtkonvexe Maximierungsaufgabe mit linearer Zielfunktion gekennzeichnet werden.

Die Dimension des Modells ergibt sich bei Berücksichtigung

eines Gasnetzes	mit 350 Knotenpunkten u. 500 Kanten,
eines Heizwassernetzes	mit 60 Knotenpunkten u. 70 Kanten,
eines Stromnetzes	mit 60 Zellengruppen

zu 3.092×4.180 , wenn 400 Zellen gebildet werden.

3. Praktische Berechnung des Modells

Für die im Modell betrachteten Netze sind bezüglich der Energieträger Gas und Heizwasser folgende Annahmen gültig [3, S. 96-100; 4] :

- a) die Strömung ist turbulent
- b) die Rohrwandungen verhalten sich wie im Übergangsbereich zwischen hydraulisch glattem und hydraulisch rauhem Verhalten.

Zur Berechnung des Druckabfalls Δp längs eines Einzelrohres, welches von einem Fluß q durchflossen wird, kann dann Gleichung (12) in folgender Form angegeben werden [3, S. A2; 4] :

$$\Delta p (q) = a q^g(q) , \quad (16)$$

worin der Faktor q neben Rohrkonstanten das spezifische Gewicht des durchfließenden Mediums beschreibt, und der Exponent $g (q)$ zunächst nicht explizit darstellbar ist [4] . Somit ist die Form der Druckabfallgleichung gemäß Gleichung (16) für eine praktische Berechnung des Modells nicht geeignet.

3.1 Approximation der Druckabfallgleichung für Einzelrohre

Die Gleichung (16) wird zur Darstellung des Druckverlustes für Einzelrohre im Modell linear approximiert. In Anbetracht der Dimension des Modells wurde eine Approximation durch lediglich zwei Geraden angestrebt. Die Zulässigkeit einer solchen Vorgehensweise konnte durch umfangreiche Testnetzberechnungen [3, S. A9 ff.] nachgewiesen werden.

Die Berechnung der approximierenden Geraden kann nach der Methode der kleinsten Quadrate erfolgen. Mit der Bestimmung der beiden Geraden hat man dann einen die Druckabfallgleichung

$\Delta p(q)$ approximierenden Polygonzug $f^*(q)$ zur Verfügung, so daß das Modell mit den Methoden der Separablen Programmierung berechnet werden kann.

Zur Darstellung des approximierenden Polygonzuges $f^*(q)$ werden die in Abb. 3 dargestellten Bezeichnungen gewählt. Führt man nun noch spezielle Variable

$$\lambda_k, \quad 0 \leq \lambda_k \leq 1 \quad (k=1, \dots, 4) \quad (17)$$

ein, so lassen sich alle Punkte des Polygonzuges $f^*(q)$ im Intervall $f_1^* \leq f^*(q) \leq f_4^*$ entsprechend der bei der λ -Methode der Separablen Programmierung üblichen Art und Weise folgender-

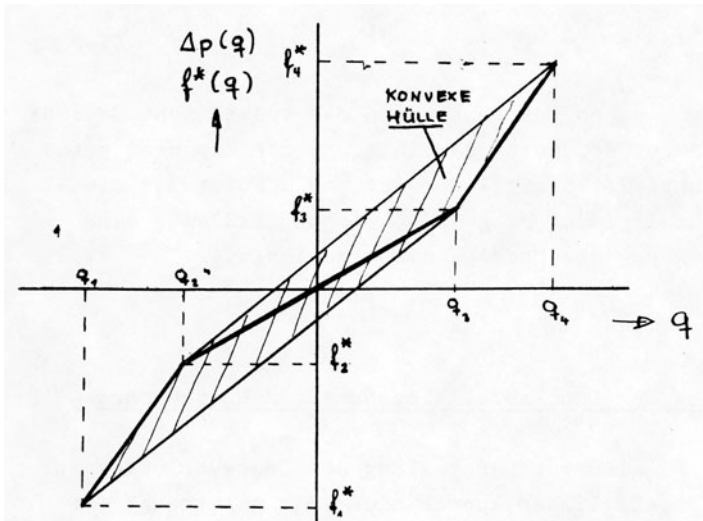


Abb. 3

Approximierender Polygonzug der Druckabfallgleichung für Einzelrohre und dessen konvexe Hülle

maßen beschreiben:

$$f^*(q) = \sum_{k=1}^4 f_k^* \lambda_k, \quad (18)$$

mit
$$\sum_{k=1}^4 \lambda_k = 1, \lambda_k \geq 0 \quad (19)$$

und a) nicht mehr als zwei λ_k positiv, (20)
 b) nur benachbarte λ_k positiv .

Für die Variable q im Intervall $q_1 \leq q \leq q_4$ gilt dann analog:

$$q = \sum_{k=1}^4 q_k \lambda_k, \quad (21)$$

mit
$$\sum_{k=1}^4 \lambda_k = 1, \lambda_k \geq 0 \quad (22)$$

und den Bedingungen (20).

Damit können jetzt die Nebenbedingungen (9) und (13) umgeformt werden zu:

$$\sum_{b=1}^4 y_{ib}^1 r_b^1 - F^1 \sum_{j=1}^{N^1} t_{sj}^1 \sum_{k=1}^4 q_{jk}^1 \lambda_{jk}^1 = -z_i^1, \quad (9a)$$

(l=1,2) ,
 (s=1,...,K¹) ,

mit 

und

$$p_s^1 - p_{s+1}^1 - t_{sj}^1 \sum_{k=1}^4 f_{jk}^{*1} \lambda_{jk}^1 = 0 \quad (l=1,2), \quad (13a)$$

(j=1,...,N¹) ,

wobei

$$\sum_{k=1}^4 \lambda_{jk}^1 = 1, \quad \lambda_{jk}^1 \geq 0$$

und die Bedingungen (2o) erfüllt sein müssen.

3.2 Bestimmung des Optimums

Da die oben formulierte Maximierungsaufgabe im allgemeinen nicht konvex ist, wird man im allgemeinen ein lokales Optimum als Lösung dieser Maximierungsaufgabe erhalten.

Es soll jetzt ein Verfahren skizziert werden, das die Bestimmung des globalen Optimums bei den zum Test des Modells verfügbaren Daten in zwei Schritten ermöglichte.

Der erste Schritt dieses Verfahrens besteht in der Berechnung einer oberen Grenze (Bound) für den Wert des globalen Optimums der zu lösenden Maximierungsaufgabe. Zu diesem Zweck wird für jede Teilstrecke $j (j=1, \dots, N^1; l=1, 2)$ die Menge P_j^1 aller Punkte des die Druckabfallgleichung approximierenden Polygonzuges $f_j^{*1} (q_j^1)$ durch deren konvexe Hülle $H_j^1 (P_j^1)$ ersetzt. Dazu genügt es hier, einfach die Bedingungen (2o), die einen beschränkten Basiseintritt bezüglich der speziellen Variablen λ_{jk}^1 fordern, fallen zu lassen. Damit erhält man dann eine "modifizierte" konvexe Maximierungsaufgabe, deren globales Optimum mit den bei der linearen Programmierung üblichen Methoden berechnet werden kann. Der Wert des so bestimmten Optimums ist außerdem Bound der ursprünglichen Maximierungsaufgabe.

Im zweiten Schritt des Verfahrens werden nun zunächst alle Flüsse $q_j^1 (j=1, \dots, N^1; l=1, 2)$ mittels der beim ersten Schritt erhaltenen Lösungswerte λ_{jko}^1 der speziellen Variablen λ_{jk}^1

gemäß Gleichung (21) berechnet. Die so berechneten Flüsse q_j^1 erfüllen alle Knotenpunktsbedingungen sowohl der "modifizierten" als auch der ursprünglichen Maximierungsaufgabe, da diese Bedingungen linear sind.

Anschließend werden zur Darstellung der eben berechneten Flüsse Werte λ_{jk}^{1*} für die speziellen Variablen λ_{jk}^1 so gesucht, daß diese die Bedingungen (20) erfüllen.

In der beim ersten Schritt gefundenen optimalen Lösung werden nun alle Werte der speziellen Variablen dieser Lösung durch λ_{jk}^1 ersetzt. Damit erhält man einen Vektor, der eine (i.a. nicht zulässige) Startbasis für die ursprüngliche Maximierungsaufgabe bildet.

Ergibt die erneute Berechnung der Maximierungsaufgabe - unter Einhaltung der Bedingungen (20) - einen Zielfunktionswert, der dem im ersten Schritt ermittelten Bound entspricht, so hat man das globale Optimum der ursprünglichen Maximierungsaufgabe gefunden.

4. Approximationsfehler

Zur Sicherstellung eines zulässigen Betriebszustandes sind die Druckbestimmungen an den Knotenpunkten mittels der in das Modell integrierten Netzberechnungen von ausschlaggebender Bedeutung.

Um die Fehler, die bezüglich der Druckbestimmung an den Knotenpunkten der Netze bei der Berechnung des Modells auftreten, abzuschätzen, sind Netzberechnungen mittels konventioneller Verfahren [1, 2, 4, 5] durchgeführt und deren Ergebnisse mit denen der in das Modell integrierten Netzberechnungen verglichen werden.

Dabei ergab sich, daß die Abweichung zwischen den im Modell bestimmten Drücken von den konventionell berechneten beim Gasnetz ($l=1$) an 94% aller Knotenpunkte nicht über 3 mm WS lagen. Diese Abweichungen sind für praktische Belange jedoch unerheblich, da sie innerhalb der Fehlergrenze praktisch realisierbarer Messungen liegen. Auch beim Heizwassernetz bewegen sich die Approximationsfehler innerhalb des für die Praxis maßgeblichen Toleranzbereiches [3, S. 162 ff] .

5. Rechenzeiten

Das oben erwähnte Zwei-Schritt-Verfahren wurde mit dem üblichen Ein-Schritt-Verfahren, d.h. die Bedingungen (2o) sind stets bindend, zur Lösung eines Testproblems verglichen. Die Dimension des hierbei verwendeten Testmodells, dem eine etwas andere Formulierung als die hier skizzierte zugrunde lag, ergab sich zu 700 Zeilen und 1267 strukturellen Variablen, wobei die Dichte der Matrix vom benutzten LP-Code zu 0,56 berechnet wurde. (Zur Lösung wurde der LP-Code OPTIMA, Vers. 3.0, benutzt, der auf einer Rechenanlage CDC 6600 zum Ablauf gebracht wurde.)

Die Ergebnisse dieses Vergleichs sind in Tabelle 1 dargestellt. Im Kopf dieser Tabelle ist die Central Processor- (CP-) Zeit, die von der Eingabe des zum Ablauf des LP-Codes notwendigen Steuerkartensatzes bis zur Ausgabe der Rechenergebnisse benötigt wurde, mit "gesamte CP-Zeit" bezeichnet.

Der Tabelle 1 kann entnommen werden, daß die Startbasis für Schritt zwei des Zwei-Schritt-Verfahrens in der Nähe einer optimalen zulässigen Lösung der ursprünglichen Maximierungsaufgabe liegt, da in Schritt zwei lediglich 8 Iterationen bis zum Erhalten des Optimums benötigt wurden.

Verfahren	Anzahl der Iterationen bis zur Lösung	Gesamte CP-Zeit [sec]	CP-Zeit bis zum Aufbau der 1.Basis [sec]	CP-Zeit zur Durchführung der Iterationen sec
<u>Ein-Schritt</u>	<u>1034</u>	<u>1334,49"</u>	<u>149,64"</u>	<u>1159,104"</u>
<u>Zwei-Schritt</u>				
Schritt eins	1282	696,33"	149,154"	521,648"
Schritt zwei	8	216,319"	157,292"	5,244"
Summe	<u>1290</u>	<u>912,649"</u>	<u>306,446"</u>	<u>526,892"</u>

Tabelle 1

Rechenzeitvergleich zwischen Ein- und Zwei-
Schritt-Verfahren

Vergleicht man die Rechenzeiten beider Verfahren - es sind die gestrichelt gekennzeichneten Werte zu vergleichen -, so ergibt sich für das Zwei-Schritt-Verfahren eine wesentlich geringere Rechenzeit.

Literaturhinweise

- [1] CIALA, H. Berechnung der Druck- und Geschwindigkeitsverhältnisse in Rohrleitungssystemen mit mehreren Einspeisungen, in: Elektrizitätswirtschaft, Bd. 64 (1965), S. 150-152
- [2] CROSS, H. Analysis of Flow in Networks of Conduits or Conductors, University of Illinois Eng. Exp. Station, Bull. 286, Nov. 1936
- [3] LIESENFELD, K.P. Ein Modell zur Absatzplanung von leitungsgebundenen Energieträgern zur optimalen Wärmeversorgung durch kommunale Querverbundunternehmen, Dissertationsschrift, TH Aachen 1970
- [4] NEUMANN, G. Elektronische Berechnung der Strömungsverhältnisse in Fernheizwassernetzen beliebiger Struktur, in: Energie, Bd. 21 (1969), S. 97-101
- [5] RENOARD, P., FLOUR, C. Fortschritte in der physikalischen und wirtschaftlichen Berechnung von Verteilungsnetzen, in: GWF (Gas), Bd. 108 (1967), S. 74

Stochastische Entscheidungsprobleme

Entscheidungen einer Person unter Unsicherheit

von R. Rabusseau und W. Reich, Bochum und Heidelberg

Vorbemerkung

Dieser Vortrag beruht auf einer Diplomarbeit von Herrn Reich und mir in Heidelberg an der mathematischen Fakultät. Die Diplomarbeit selbst ist die Ausarbeitung und Ergänzung der von Herrn Dr. W. Böge im Sommer-Semester 1969 gehaltenen Vorlesung "Spieltheorie".

Ich möchte Sie daher bitten, Herrn Reich und mich als Stellvertreter von Herrn Dr. Böge anzusehen, mit dessen Einwilligung wir hier vortragen.

Es handelt sich um die axiomatische Kennzeichnung konsistenter Strategie-Auswahlprinzipien und die Herleitung der Existenz einer persönlichen reellen Nutzenfunktion und deren Eindeutigkeit bis auf isotone affine Transformationen, sowie der Existenz und Eindeutigkeit einer persönlichen Wahrscheinlichkeitsverteilung.

Die Arbeit gehört in die BAYESsche Entscheidungstheorie, d.h. sie benutzt subjektive Wahrscheinlichkeiten und wurde angeregt durch das Buch von L.J. SAVAGE: The Foundations of Statistics. Allerdings werden die Axiome von L.J. SAVAGE teilweise von W. Böge durch andere ersetzt. Anders als bei SAVAGE wurde nicht zuerst die Existenz einer persönlichen Wahrscheinlichkeitsverteilung, sondern zunächst aus den Axiomen (S K), (R K) und (A T) die Existenz und Eindeutigkeit bis auf isotone affine Transformationen eines bezüglich eines BOOLEschen Verbandes der Ereignisse additiven reellen Funktionals auf der Menge der Strategien, welches mit den Präferenzen verträglich ist, gefolgert. Aus weiteren Axiomen, nämlich (E) und (Z) wurde dann gefolgert, daß dieses Funktional ein Integral bezüglich eines eindeutig bestimmten Wahrscheinlichkeitsmaßes ist.

Beispiel:

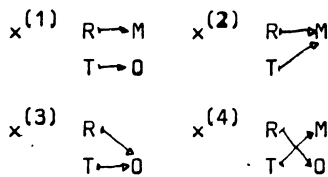
P sei gezwungen, das Haus zu verlassen und habe die möglichen Handlungen mit M oder ohne O Regenschirm zu gehen.

$$H = \{M, O\}$$

Die möglichen externen Schicksale seien Regenwetter R oder trockenes Wetter T

$$S = \{R, T\}$$

$X := H^S$ besteht aus folgenden Strategien (Abbildungen):



Es ergeben sich die Konsequenzen:

c	M	O	
R	t	n	t trocken
T	p	t	n naß
			p peinlich

$x^{(1)*} \quad R \mapsto t$ $\quad \quad T \nearrow$	$x^{(2)*} \quad R \mapsto t$ $\quad \quad T \mapsto p$
$x^{(3)*} \quad R \mapsto n$ $\quad \quad T \mapsto t$	$x^{(4)*} \quad R \mapsto n$ $\quad \quad T \mapsto p$

Zur Beschreibung von P würde genügen, wie P in einer gegebenen Situation aus der Menge X ihrer möglichen Strategien diejenigen aussucht, die sie tatsächlich anwendet.

Wir nehmen aus Gründen der mathematischen Einfachheit an, daß P in jeder konkreten Situation nur aus einer endlichen Teilmenge Y von X auswählen kann. Sei $\text{Pot}_e X$ die Menge der endlichen Teilmengen von X . Diese Auswahl denken wir uns gegeben durch eine Abbildung (Strategieauswahlprinzip)

$$f : \text{Pot}_e X \longrightarrow \text{Pot}_e X$$

Wir fordern, daß f dem sogenannten Schnittkonsistenzaxiom (SK) genügt:

$$\begin{aligned} \text{(i)} \quad & \bigwedge_{M \in \text{Pot}_e X} f M \subseteq M \\ \text{(ii)} \quad & \bigwedge_{\emptyset \neq M \in \text{Pot}_e X} f M \neq \emptyset \\ \text{(SK)} \quad \text{(iii)} \quad & \bigwedge_{M, N \in \text{Pot}_e X} [N \cap f M \neq \emptyset \implies f(N \cap M) = N \cap f M] \end{aligned}$$

Sei \mathcal{F} die Menge der (SK)-Auswahlprinzipien, \mathcal{Q} die Menge der Quasi-anordnungen (auch Rangordnungen) auf (in) X (d.h. vergleichbar + transitiv). Dann gilt der

SATZ Die Quasianordnungen einer Menge X und die (SK)-Strategieauswahlprinzipien auf X entsprechen sich umkehrbar eindeutig vermöge

$$\begin{aligned} \mathcal{Q} \ni \leq & \longmapsto f \\ \bigwedge_{M \in \text{Pot}_e X} f M & := \{x \in M, \bigwedge_{y \in M} y \leq x\} \\ f \ni f & \longmapsto \leq \\ \bigwedge_{x, y \in X} y \leq x & \xleftrightarrow{\text{Def.}} x \in f\{x, y\} \end{aligned}$$

Wir gehen daher im folgenden immer von einer (festen) Quasianordnung auf X aus.

Bem.: Sei $X := \{x_1, \dots, x_n\}$ eine endliche Menge und \preceq eine Quasianordnung in X . Dann gilt

\preceq ist durch $3/2 n$ Angaben bestimmt (im Gegensatz zu n^2)

Für die Anzahl der Quasianordnungen $S(n)$ in der Menge X ergibt sich

$$S(n) = \sum_{k=1}^n S(n, k), \text{ wobei}$$

$$S(n, k) := \sum_{n_1 + \dots + n_k = n} \frac{n!}{n_1! \dots n_k!}$$

$$1 \leq n_1 \leq n - k + 1$$

Aufgrund der Rekursionsformel:

$$S(n, k) = k \cdot [S(n-1, k-1) + S(n-1, k)]$$

wurden die $S(n, k)$ und $S(n)$ mit dem Computer errechnet. Es ergaben sich die folgenden Tabellen.

$S(1,1)$	$S(2,k)$	$S(3,k)$	$S(4,k)$	$S(5,k)$
1	1	1	1	1
	2	6	14	30
		6	36	150
			24	240
				120
<hr/>				
1	3	13	75	541

S (6, k)	S (7, k)	S (8, k)	S (9, k)	S (10, k)
1	1	1	1	1
62	126	254	510	1 022
540	1 806	5 796	18 150	55 980
1 560	8 400	40 824	186 480	818 520
1 800	16 800	126 000	834 120	5 103 000
720	15 120	191 520	1 905 120	16 435 440
	5 040	141 120	2 328 480	29 635 200
		40 320	1 451 520	30 240 000
			362 880	16 329 600
				3 628 800

4 683	47 293	545 835	7 087 261	102 247 563
-------	--------	---------	-----------	-------------

S (11, k)	S (12, k)	S (13, k)	S (14, k)
1	1	1	1
2 046	4 094	8 190	16 382
171 006	519 156	1 569 750	4 733 820
3 498 000	14 676 024	60 780 720	249 401 880
29 607 600	165 528 000	901 020 120	4 809 004 200
129 230 640	953 029 440	6 711 344 640	45 674 188 560
322 494 480	3 162 075 840	28 805 736 960	248 619 571 200
479 001 600	6 411 988 640	76 592 355 840	843 184 742 400
419 126 400	8 083 152 000	130 456 085 760	1 863 435 974 400
199 584 000	6 187 104 000	142 702 560 000	2 731 586 457 600
39 916 800	2 634 508 800	97 037 740 800	2 637 143 308 800
	479 001 600	37 362 124 800	1 612 798 387 200
		6 227 020 800	566 658 892 800
			87 178 291 200

1 622 632 573	28 091 567 595	526 858 348 381	10 641 342 970 443
---------------	----------------	-----------------	--------------------

S (15, k)	S (16, k)	S (17, k)
1	1	1
32 766	65 534	131 070
14 250 606	42 850 116	128 746 950
1 016 542 800	4 123 173 624	16 664 094 960
25 292 030 400	131 542 866 000	678 330 198 120
302 899 156 560	1 969 147 121 760	12 604 139 926 560
2 060 056 318 320	16 540 688 324 160	129 568 848 121 440
8 734 434 508 800	86 355 926 616 960	823 172 919 528 960
24 359 586 451 200	297 846 188 640 000	3 457 819 037 312 640
45 950 224 320 000	703 098 107 712 000	10 009 442 963 520 000
59 056 027 430 400	1 155 068 769 254 400	20 439 835 646 630 400
50 999 300 352 000	1 320 663 933 388 800	29 708 792 431 718 400
28 332 944 640 000	1 031 319 184 896 000	30 575 780 537 702 400
9 153 720 576 000	524 813 313 024 000	21 785 854 970 880 000
1 307 674 368 000	156 920 924 160 000	10 226 013 557 760 000
	20 922 789 888 000	2 845 499 424 768 000
		355 687 428 096 000
230 283 190 977 853	5 315 654 681 981 355	130 370 767 029 135 901

S (18, k)	S (19, k)	S (20, k)
1	1	1
262 142	524 286	1 048 574
386 634 060	1 160 688 606	3 483 638 676
67 171 367 640	270 232 006 800	1 085 570 781 624
3 474 971 465 400	17 710 714 165 200	89 904 730 860 000
79 694 820 748 080	499 018 753 280 880	3 100 376 804 676 480
995 210 916 336 000	7 524 340 159 588 560	56 163 512 390 086 080
7 621 934 141 203 200	68 937 160 460 313 600	611 692 004 959 217 280
38 528 927 611 574 400	415 357 755 774 998 400	4 358 654 246 117 808 000
134 672 620 008 326 400	1 732 015 476 199 008 000	21 473 732 319 740 064 000
334 942 064 711 654 400	5 165 761 531 919 788 800	75 875 547 089 306 764 800
601 783 536 940 185 600	11 240 707 219 822 080 000	196 877 625 020 902 425 600
783 699 448 602 470 400	18 011 278 812 054 528 000	380 275 818 414 395 904 000
733 062 897 120 153 600	21 234 672 840 116 736 000	549 443 323 130 397 696 000
480 178 027 929 600 000	18 198 613 875 746 304 000	591 499 300 737 945 600 000
209 144 207 720 448 000	11 029 155 770 400 768 000	467 644 314 338 353 152 000
54 420 176 498 688 000	4 480 594 531 725 312 000	263 665 755 136 143 360 000
6 402 373 705 728 000	1 094 805 903 679 488 000	100 357 207 837 286 400 000
	121 645 100 408 832 000	23 112 569 077 678 080 000
		2 432 902 008 176 640 000
3 385 534 663 256 845 323	92 801 587 319 328 411 133	2 677 687 796 244 384 203 115

Def.: $A \subseteq S$ heißt Zerlegungsmenge für X , falls eine der äquivalenten Bedingungen gilt:

- (i) $\bigwedge_{x, y \in X} x_A * y_{S \setminus A} \in X$
- (ii) $X_A * X_{S \setminus A} \subseteq X$
- (iii) $X_A * X_{S \setminus A} = X$

Erklärung: $x : S \longrightarrow H, x_A : A \longrightarrow H$

$$x_A(a) = x(a)$$

x_A ist die Restriktion von x auf A .

$$X_A := \{x_A, x \in X\}, (x_A * y_{S \setminus A})_A := x_A$$

$$(x_A * y_{S \setminus A})_{S \setminus A} := y_{S \setminus A}$$

SATZ Das System $\mathcal{L}(X) := \{A \subseteq S, X_A * X_{S \setminus A} = X\}$ der Zerlegungsmengen von X ist eine BOOLEsche Algebra, d.h.

$$A, B \in \mathcal{L}(X) \implies A \cup B \in \mathcal{L}(X), S \in \mathcal{L}(X)$$

$$A \in \mathcal{L}(X) \implies S \setminus A \in \mathcal{L}(X)$$

Sei \mathcal{L} eine BOOLEsche Unter algebra von $\mathcal{L}(X)$.

2. DAS RESTRIKTIONSKONSISTENZ-AXIOM (RK)

Das Axiom (RK) lautet:

Für alle $A \in \mathcal{L}$, alle $u, u' \in X_A$, alle $v, v' \in X_{S \setminus A}$ gilt:
(RK)

$$u * v \succeq u' * v \iff u * v' \succeq u' * v'$$

Def.: Für $A, B \in \mathcal{L}$ sei

$$A \succeq_{x \rightarrow y} B \iff y_A * x_{S \setminus A} \succeq y_B * x_{S \setminus B} \quad \text{Def.}$$

Wir sagen: A ist nicht schlechter als B für den Übergang von x nach y .

Bem.: $\succeq_{x \rightarrow y}$ ist eine Quasianordnung auf \mathcal{L} , die die folgende Bedingung der Additivität erfüllt:

$$\left. \begin{array}{l} A \succeq_{x \rightarrow y} B, C \succeq_{x \rightarrow y} D \\ A \cap C = B \cap D = \emptyset \end{array} \right\} \implies A \cup C \succeq_{x \rightarrow y} B \cup D$$

Def.: Ein System $\{A_1, \dots, A_n\}$ von Teilen von S heißt Zerlegung von S , falls gilt:

$$(i) \quad \bigcup_{i=1}^n A_i = S$$

$$(ii) \quad \bigwedge_{i \neq j} A_i \cap A_j = \emptyset$$

Wir zeigen anhand eines Beispielles, daß nicht jede additive Quasianordnung $\succeq_{x \rightarrow y}$ auf \mathcal{L} durch ein Maß ν erzeugt wird, d.h.

$$\bigwedge_{A, B \in \mathcal{L}} A \succeq_{x \rightarrow y} B \iff \nu(A) \geq \nu(B)$$

Man muß allerdings für S eine mindestens 5-elementige Menge wählen¹⁾

1) vgl. C. KRAFT, J. PRATT, A. SEIDENBERG: Intuitive Probability on Finite Sets. Annals of Math. Statistics 1959, 30, S. 408-419,
A. KIRSCH: Gerechte lineare Ordnungen und Punktbewertungen, in: Der Mathematikunterricht 1969/1, S. 64-84.

Aus (RK) folgt das

Vereinigungslemma (VL): Seien $x, y \in X$, $A_1, \dots, A_n, B_1, \dots, B_n \in \mathcal{L}$ mit $A_i \gtrsim_{x \rightarrow y} B_i$, $i = 1, \dots, n$. $A_i \cap A_j = B_i \cap B_j = \emptyset$

für $i \neq j$. Dann gilt:

$$1.) \quad \bigcup_1^n A_i \gtrsim_{x \rightarrow y} \bigcup_1^n B_i$$

2.) Es gebe ein $i \in \{1, \dots, n\}$ mit $A_i >_{x \rightarrow y} B_i$. Dann gilt

$$\bigcup_1^n A_i >_{x \rightarrow y} \bigcup_1^n B_i$$

Bem.: (RK) \Longleftrightarrow (VL)

Divisionslemma (DL): Seien $m, n \in \mathbb{N}^*$, $x, y \in X$

$$\emptyset = A_0, A_1, \dots, A_m, A_i \cap A_j = \emptyset \text{ für } i \neq j$$

$$\emptyset = B_0, B_1, \dots, B_n, B_i \cap B_j = \emptyset \text{ für } i \neq j$$

seien Mengen aus \mathcal{L} mit

$$\emptyset = A_0 \lesssim_{x \rightarrow y} A_1 \lesssim_{x \rightarrow y} \dots \lesssim_{x \rightarrow y} A_m$$

$$\emptyset = B_0 \lesssim_{x \rightarrow y} B_1 \lesssim_{x \rightarrow y} \dots \lesssim_{x \rightarrow y} B_n \quad \text{und}$$

$$\bigcup_1^n A_i \lesssim_{x \rightarrow y} \bigcup_1^n B_i$$

Dann gilt:

$$\bigwedge_{r=0}^m \bigwedge_{s=0}^n \left[\frac{r}{m} \leq \frac{s}{n} \Rightarrow \bigcup_1^r A_i \lesssim_{x \rightarrow y} \bigcup_{i=n-s+1}^n B_i \right]$$

3. DAS ARCHIMEDISCHE UND TEILUNGSAXIOM (AT)

Axiom (\tilde{AT}) Zu $x, y \in X$, $A \in \mathcal{L}$, $A \succ_{x \rightarrow y} \emptyset$ gibt es eine Zerlegung

$$\{A_1, \dots, A_n\} \subseteq \mathcal{L} \text{ von } S, n \in \mathbb{N}^*, \text{ mit } A_i \prec_{x \rightarrow y} A \text{ f\"ur } i = 1, \dots, n.$$

Mit Hilfe der bis jetzt formulierten Axiome kommt man fast bis zum ersten Hauptsatz. Nur an einer Stelle ben\"otigt man das folgende st\"arkere

Axiom (AT) Zu $A \succ_{x \rightarrow y} \emptyset$, $B \in \mathcal{L}$ gibt es eine Zerlegung

$$\{B_1, \dots, B_n\} \subseteq \mathcal{L} \text{ von } B \text{ mit } B_i \prec_{x \rightarrow y} A \text{ f\"ur } i = 1, \dots, n.$$

F\"ur $B := S$ erh\"alt man (\tilde{AT}). Wir vermuten, da\B aus (\tilde{AT}) auch (AT) folgt, sind aber bis jetzt nicht in der Lage, dies zu beweisen.

Def.: Das Tripel $(X, \succeq, \mathcal{L})$ hei\Bt (SB)-Raum (SAVAGE-B\"OGE-Raum), falls \succeq Quasianordnung auf X ist, die den Axiomen (RK) und (AT) gen\"ugt.

In (\tilde{AT}) bzw. (AT) werden f\"ur den Fall, da\B es $x < y$ gibt, zum ersten Mal Existenzen gefordert.

Das Axiom (AT) mag zun\"achst recht k\"unstlich aussehen. Denn w\"ahrend (SK) und (RK) notwendig f\"ur die Existenz einer Nutzenfunktion sind, ist das bei (AT) nicht der Fall. Wir zeigen in der Arbeit, da\B das Axiom (AT) im Fall des erweiterten Regenschirmbeispiels (Erweiterung durch W\"urfelw\"urfe) realisiert werden kann, und zwar auf realistische Weise. Dr. B\"oge stellte die These auf, da\B diese Erweiterung auch in jedem anderen Beispiel m\"oglich ist.

4. DER 1. HAUPTSATZ

Def.: Eine Abbildung $N : X \rightarrow \mathbb{R}$ heißt \mathcal{L} -additiv, wenn es eine Familie $(N_A)_{A \in \mathcal{L}}$ von Abbildungen $N_A : X_A \rightarrow \mathbb{R}$ gibt, so daß für jede Zerlegung $\{A_1, \dots, A_n\} \subseteq \mathcal{L}$ von S gilt:

$$N(x) = \sum_{i=1}^n N_{A_i}(x_{A_i}) \quad \forall x \in X$$

Def.: Eine Abbildung $N : X \rightarrow \mathbb{R}$ heißt (persönl.) Nutzenfunktion von P , falls

1. N ist \mathcal{L} -additiv
2. $\bigwedge_{x, y \in X} x \succeq y \Rightarrow N(x) \geq N(y)$

Wir geben nun 3 Formulierungen des 1. Hauptsatzes.

1. Hauptsatz^{*} : Sei $(X, \succeq, \mathcal{L})$ ein (SB)-Raum. Dann gilt:

(i) Es gibt eine Nutzenfunktion N auf X

(ii) Falls $S \neq \emptyset$ ist, so gilt:

\tilde{N} ist eine weitere Nutzenfunktion auf X \Leftrightarrow

$$\bigvee_{\substack{a > 0 \\ b \in \mathbb{R}}} \tilde{N} = a \cdot N + b$$

Zusatz: In (ii) gilt " \Rightarrow " auch ohne die Vor. $S \neq \emptyset$.

1. Hauptsatz^{**} : Sei $(X, \succeq, \mathcal{L})$ ein (SB)-Raum. Dann gilt:

(i) Es gibt eine Nutzenfunktion N auf X

(ii) \tilde{N} ist eine weitere Nutzenfunktion auf X mit \mathcal{L} -additiver Zerlegung $(\tilde{N}_A)_{A \in \mathcal{L}}$ \Leftrightarrow

$$\bigvee_{a > 0} \bigvee_{\substack{\text{endl. additives} \\ \text{Maß } \beta: \mathcal{L} \rightarrow \mathbb{R}}} \tilde{N}_A = a \cdot N_A + \beta(A)$$

Def.: Eine Abbildung $\tilde{N} : X \rightarrow \mathbb{R}$ heißt eine Pseudonutzenfunktion auf X , falls gilt

(i) \tilde{N} ist \mathcal{L} -additiv

(ii) $\bigwedge_{x, y \in X} [x \succeq y \Rightarrow N(x) \geq N(y)]$

1. Hauptsatz: Sei $(X, \succeq, \mathcal{L})$ ein (SB)-Raum.

(i) Dann gibt es eine Nutzenfunktion N auf X

(ii) Falls $S \neq \emptyset$ ist, so gilt:

\tilde{N} ist eine Pseudonutzenfunktion auf $X \iff$
 $\forall \begin{matrix} a \geq 0 \\ b \in \mathbb{R} \end{matrix} \quad \tilde{N} = a \cdot N + b$

5. DER 2. HAUPTSATZ

Def.: Die Konsequenzfunktion $c : S \times H \rightarrow C$ heißt zulässig (für \succeq), falls eine der beiden äqu. Bed. gilt:

$$(i) \quad x^* = y^* \Rightarrow x \sim y$$

$$(ii) \quad \text{Es gibt eine Quasianordnung } \succeq^* \text{ auf } X^* \text{ mit}$$

$$x \succeq y \Leftrightarrow x^* \succeq^* y^*$$

Def.: $B \in \mathcal{L}_A$ heißt präferenzirrelevant für $Y \subseteq X_A$, wenn für alle $x, y \in Y$ gilt:

$$x \succeq_A y \Rightarrow x_B \succeq y_B$$

Def.: $B \in \mathcal{L}_A$ heiße vollständig präferenzirrelevant für $Y \subseteq X_A$, wenn es präferenzirrelevant für eine geeignete Erweiterung Y' ist ($Y \subseteq Y' \subseteq X_A$) die bezüglich eines geeigneten BOOLEschen Verbandes $\mathcal{L}' \subseteq \mathcal{L}_A \cap \mathcal{L}(Y')$ das Axiom (AT) erfüllt.

Def.: Die Konsequenzfunktion $c : S \times H \rightarrow C$ heißt erschöpfend, wenn sie zulässig ist und wenn jedes $B \in \mathcal{L}_A$ vollständig präferenzirrelevant ist für jede endliche Menge konstanter Konsequenzstrategien in X_A , für alle $A \in \mathcal{L}$

Axiom (E) Die Konsequenzfunktion c sei erschöpfend.

Wir formulieren nun das letzte Axiom (Z). In ihm werden Existenzen gefordert.

Def.: Ein System \mathcal{L} von Teilen t einer gegebenen Menge heißt abstrakter simplizialer Komplex, falls gilt:

$$(i) \quad \bigwedge_{t \in \mathcal{L}} [t' \subseteq t \Rightarrow t' \in \mathcal{L}]$$

$$(ii) \quad \text{Jedes } t \in \mathcal{L} \text{ ist ein endl. Teil der gegebenen Menge.}$$

Beispiele von simplizialen Komplexen:

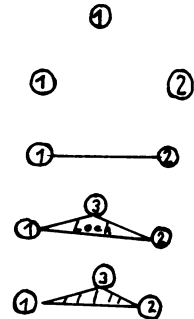
$$\mathcal{L}_1 : = \{\emptyset, \{1\}\}$$

$$\mathcal{L}_2 : = \{\emptyset, \{1\}, \{2\}\}$$

$$\mathcal{L}_3 : = \text{Pot } \{1, 2\}$$

$$\mathcal{L}_4 : = (\text{Pot } \{1, 2, 3\}) \setminus \{1, 2, 3\}$$

$$\mathcal{L}_5 : = \text{Pot } \{1, 2, 3\}$$



Zusammenhangsaxiom (Z):

Es gibt eine Zerlegung $\{D_1, \dots, D_m\} \subseteq \mathcal{L}$ von S und eine Überdeckung $C = C_1 \cup \dots \cup C_m$ mit $C_i \neq \emptyset$, $i = 1, \dots, m$, so daß

1. $X_{D_1}^*$ = Menge aller \mathcal{L}_{D_1} -Treppenfunktionen mit Werten in C_1

2. $C_{ijk} := C_i \cap C_j \cap C_k \neq \emptyset \implies$

$$\exists \gamma, \gamma' \in C_{ijk} \text{ mit } \gamma_{D_v} <^*_{D_v} \gamma'_{D_v}, \quad v \in \{i, j, k\}$$

3. Der abstrakte simpliziale Komplex \mathcal{L} mit

$$\mathcal{L} := \{t \subseteq \{1, \dots, m\}, \bigcap_{i \in t} C_i \neq \emptyset\}$$

ist ⁴zusammenhängend und einfach zusammenhängend.

2. Hauptsatz: Sei $\mathcal{X} := (X, \succeq, \mathcal{L})$ ein (SB)-Raum, der auch die Axiome (E) und (Z) erfüllt. Dann gibt es eine (ebenfalls mit N bezeichnete) Nutzenfunktion

$$N : X^* \longrightarrow \mathbb{R}, \text{ die zu der Nutzenfunktion}$$

$$N : X \longrightarrow \mathbb{R} \quad \text{in der Beziehung steht:}$$

$$N(x) = N(x^*) \quad \forall x \in X$$

Weiter gibt es ein endlich additives subjektives Wahrscheinlichkeitsmaß

$$P : \mathcal{L} \longrightarrow [0, 1] \quad \text{und eine subjektive Nutzenfunktion}$$

$$N^* : C \longrightarrow \mathbb{R}, \text{ so daß}$$

$$N = \int N^* dP, \text{ d.h.}$$

$$\bigwedge_{x^* \in X^*} N(x^*) = \sum_1^n N^*(c_i) P(A_i), \text{ wobei}$$

$\{A_1, \dots, A_n\} \in \mathcal{L}$ irgendeine Zerlegung von S ist mit

$$x^*(A_i) \subseteq \{c_i\} \text{ (solche Zerlegungen gibt es auch).}$$

Das Auswahlprinzip f der Person ist bestimmt durch

$$f(Y) = \left\{ x \in Y, \bigwedge_{y \in Y} N(x^*) \geq N(y^*) \text{ für } Y \in \text{Pot}_0 X. \right\}$$

Durch f ist P eindeutig und N^* eindeutig bis auf einen positiven konstanten Faktor und eine additive Konstante bestimmt

d. h. sei $N = \int N^* dP$. Dann gilt

$$N' = \int N^{*'} dP \text{ mit } N' \in \mathcal{X}^* := \text{Menge der Nutzenfunktion auf}$$

$\mathcal{X} := (X, \succeq, \mathcal{L})$ genau dann, falls

$$\bigvee_{\substack{a > 0 \\ b \in \mathbb{R}}} N^{*'} = a \cdot N^* + b$$

$N(x^*)$ ist also gleich dem Erwartungswert der subjektiven Nutzenfunktion bezogen auf die Konsequenzstrategie x^* bezüglich des subjektiven Wahrscheinlichkeitsmaßes P . Wir nennen diesen Erwartungswert auch den mittleren persönlichen Nutzen von x^* .

Der 2. Hauptsatz besagt also: Der Entscheidende wählt aus jeder endlichen Menge von Handlungsstrategien diejenige aus, deren mittlerer persönlicher Nutzen maximal ist (Bernoulli-Prinzip).

Er handelt so, als würden die externen Schicksale einer Menge $A \in \mathcal{L}$ mit der objektiveren Wahrscheinlichkeit $R(A)$ eintreten, obgleich völlige Unsicherheit hinsichtlich des Eintretens der externen Schicksale besteht.

Der Entscheidende ist also in der Lage, aufgrund seiner Reflexionen über sein eigenes Auswahlprinzip f eine subjektive Wahrscheinlichkeit $P(A)$, die mit f in Einklang steht, für alle Mengen $A \in \mathcal{L}$ festzulegen.

6. PRAKTISCHE BEDEUTUNG DER ARBEIT FÜR DIE ENTSCHEIDUNGSTHEORIE

Das Ziel der Arbeit ist die Herleitung des BERNOLLI'schen Prinzips für subjektive Wahrscheinlichkeiten.

Darüber hinaus möchte die Arbeit nicht nur dem Leser ein Kriterium in die Hand geben, um Entscheidungsprobleme zu lösen, sondern auch ein Hilfsmittel, um einer zweiten Person bei der Festlegung der Präferenzordnung ihrer Strategien zu helfen. Damit werden wir in der Lage sein, für eine zweite Person, die die Strategien anwenden soll, die notwendigen Entscheidungen zu treffen. (Das benötigt man in der typischen beratenden Situation des Mathematikers im Beruf.)

Um die Präferenzordnung der Strategien angeben zu können, brauchen wir von der Person Angaben. Im Abschnitt 7 werden wir an einem Beispiel sehen, wie stark die Anzahl der Angaben zurückgeht, wenn wir die Ergebnisse der Arbeit benutzen. Mit dem 2. Hauptsatz, dem BERNOLLI-Prinzip für subjektive Wahrscheinlichkeiten, haben wir dann ein Hilfsmittel, das uns die Beschreibung der Präferenzordnung mit relativ wenig Angaben ermöglicht.

Zu dem in der Literatur weit verbreiteten noch weniger Angaben benötigenden Minimax-Prinzip werden wir ein Beispiel geben, das zu einem nicht plausiblen Resultat führen wird. Mit dem 2. Hauptsatz jedoch werden wir im selben Beispiel ein Resultat erhalten, das der Leser auch ohne mathematische Überlegungen erwartet hätte.

Somit ist die Arbeit mit dem Minimax-Prinzip nicht allgemein verträglich und möchte auch keine Modifikation desselben sein.

Bayesian Reliability Demonstration

by G. J. Schick and T. M. Drnas, Los Angeles and Culver City

ABSTRACT

A Bayesian approach to reliability demonstration testing is described and differences between the Bayesian viewpoint and the commonly employed classical approach are highlighted. A procedure for selecting a specific inverted gamma probability density to characterize the prior distribution of the MTBF of electronic hardware is developed and a table of Bayesian demonstration plans for a practical range of input parameters is provided. In addition, procedures for implementation of the plans and two illustrative examples are given. Finally, two commonly employed classical plans are compared to a Bayesian plan illustrating the efficiency of the latter in terms of demonstration test time requirements.

INTRODUCTION

Recent publications [1, 2, 3, 4, 5, 6, 10]¹ have shown that the aerospace industry is becoming increasingly interested in the advantages generated when reliability demonstration tests of electronic systems are based on Bayesian concepts as opposed to classical concepts. In particular, the major advantages include:

- A. Less test time, generally, is required to demonstrate a reliability requirement, thus, at times, reducing costs dramatically.
- B. The Bayesian approach takes advantage of prior knowledge.
- C. The Bayesian interpretation of test results appears to be more appealing to engineers.

These advantages have been discussed in considerable detail in the reference literature.

The objective of this paper is to describe a Bayesian approach to reliability demonstration, provide a table of Bayesian based demonstration plans, and specify a procedure for implementing the plans.

From the statistical viewpoint, reliability demonstration tests are sampling schemes wherein the hardware or system is tested by simulating mission conditions in order to determine in a probabilistic sense if the reliability requirement has been achieved. Thereafter, the resulting test data (sample) and preselected decision rule permit an accept or reject decision. Prior to the consideration of the Bayesian approach, reliability demonstration test decision rules have been largely based on the Neyman-Pearson theory of hypothesis testing. In many cases a demonstration test is specified in terms of total operating time, T , that the system must experience, and a maximum allowable number of failures, a . The elements T and a are determined by the following preselected factors:

- A. Underlying statistical model in the form of a probability density function of θ .
- B. Minimum acceptable MTBF, θ_1 .
- C. Specified MTBF, θ_0 , where $\theta_1 < \theta_0$.
- D. Consumer's risk, γ_2 . The probability of accepting systems with a true MTBF equal to the minimum acceptable MTBF, θ_1 . The probability of accepting systems with a true MTBF less than θ_1 will be less than or equal to γ_2 .
- E. Producer's risk, γ_1 . The probability of rejecting systems with a true MTBF equal to the specified MTBF, θ_0 . The probability of rejecting systems with a true MTBF greater than θ_0 will be less than or equal to γ_1 .²

¹ Numerals in brackets refer to References at end of paper.

² All definitions given here are in agreement with MIL-STD-781B[11].

Subsequently, minimum T and the associated acceptable number of failures, a , are determined such that they satisfy the inequalities

$$\Pr(\text{Acceptance} \mid \theta = \theta_1) = \Pr(f \leq a \mid \theta = \theta_1) \leq \gamma_2 \quad (1)$$

and

$$\Pr(\text{Rejection} \mid \theta = \theta_0) = \Pr(f > a \mid \theta = \theta_0) \leq \gamma_1 \quad (2)$$

where the probabilities are determined by operating on the probability density function and f represents the number of failures observed. For example, if the applicable statistical model is assumed to be the Poisson distribution, which is quite common in practice, the above inequalities would be written

$$\Pr(f \leq a \mid \theta = \theta_1) = \sum_{f=0}^a e^{-T/\theta_1} (T/\theta_1)^f / f! \leq \gamma_2 \quad (3)$$

and

$$\Pr(f > a \mid \theta = \theta_0) = \sum_{f=a+1}^{\infty} e^{-T/\theta_0} (T/\theta_0)^f / f! \leq \gamma_1 \quad (4)$$

Minimum T and associated a are obtained by simultaneous solution of (3) and (4). Upon concluding the test, the customer considers the MTBF requirement to be satisfied and accepts the system if the number of failures in T hours is less than or equal to a . Otherwise the system is rejected.

Wald's [12] sequential decision approach wherein a third decision, i.e., to continue testing if no accept or reject decision can be made, has reduced the expected test time to reach a decision when compared to the method just described. Wald's theory eventually resulted in the well known MIL-STD-781 (AGREE) [11] demonstration test plans now imposed on the aerospace industry in all serious reliability programs. However, due to increasing reliability requirements and the increasing complexity of aerospace hardware, the test time needed to demonstrate reliability has reached a point where the cost of such tests has become prohibitive in spite of the efficiencies of AGREE test plans. Thus, the interest in Bayesian approach and its associated advantages listed in the introductory paragraph.

BAYES THEOREM

Let H_1, H_2, \dots, H_n represent a mutually exclusive and exhaustive collection of hypotheses. Suppose that an event S exists and the conditional probabilities $P(S/H_i)$ are known. Also suppose that the probabilities $P(H_i)$ are known. $P(H_i)$ is termed the prior probability that H_i is true. Then for the discrete case the following conditional relationship is known as Bayes' Theorem:

$$P(H_i \mid S) = \frac{P(H_i) P(S \mid H_i)}{P(S)}, P(S) \neq 0$$

where

$$P(S) = \sum_{i=1}^n P(H_i) P(S|H_i)$$

$P(H_i|S)$ is the posterior probability that H_i is true upon observing a sample statistic, S . In the continuous case Bayes' Theorem may be expressed as

$$P(\theta|S) = \frac{g(\theta) h(S|\theta)}{\int_{-\infty}^{\infty} g(\theta) h(S|\theta) d\theta}$$

where $g(\theta)$ is the prior probability density function with the MTBF θ , as the random variable, $P(\theta|S)$ is the posterior probability density function with θ as the random variable and $h(S|\theta)$ represents the conditional sampling distribution of S . The theorem, as expressed above, permits probabilistic statements regarding θ upon observing the statistic S . Thus, observed information (test data) is combined with the prior density via the sampling distribution to form the posterior density. The conditional sampling distribution in Bayes' Theorem having the same interpretation as the sampling distribution in the classical approach presents no new problem. On the other hand, the prior density, represented by the density, $g(\theta)$, in Bayes' Theorem does require special attention.

RELIABILITY DEMONSTRATION-BAYESIAN APPROACH

Insofar as reliability demonstration is concerned, within the Bayesian context, the consumer is interested in being assured that the product received, i.e., the accepted product, has an MTBF no less than a minimum acceptable value, θ_1 . On the other hand, it is in the best interest of the producer that the accept-reject decision criteria be structured to assure that the true MTBF of the product is indeed less than the specified θ_0 if the product is rejected. Here, $\theta_1 < \theta_0$. Stated slightly differently, the consumer requires the probability to be $1 - \beta^*$ that the true MTBF of the product be greater than θ_1 given that the product has been accepted by the reliability demonstration test. Concurrently, the producer requires that the probability be $1 - \alpha^*$ that the true MTBF of the product be less than the specified value, θ_0 , given the product has been rejected by the reliability demonstration test. Symbolically this is equivalent to

$$\Pr(\theta \geq \theta_1 | \text{Acceptance}) \geq 1 - \beta^*$$

and

$$\Pr(\theta \leq \theta_0 | \text{Rejection}) \leq 1 - \alpha^*$$

Note that the preceding inequalities refer to probabilistic statements regarding θ after an accept or reject decision has been made.

The foregoing may be reduced to a set of mathematical inequalities in view of Bayes' Theorem. Let the prior distribution of the MTBF be known, agreed upon by the consumer and producer and represented by the density $g(\theta)$. The interpretation of the conditional sampling distribution $h(S|\theta)$ is identical to that of the sampling distribution employed in the classical approach. I.e., S is an observed statistic being a function of f failures, T test time, and θ is the MTBF. Then the requirements of the consumer and the producer may be written as follows:

$$\Pr(\theta \leq \theta_1 | f \leq a, T) = \frac{\int_0^{\theta_1} g(\theta) \sum_{f=0}^a h[f, T|\theta] d\theta}{\int_0^{\infty} g(\theta) \sum_{f=0}^a h[f, T|\theta] d\theta} \leq \beta^* \quad (5)$$

$$\Pr(\theta \leq \theta_0 | f > a, T) = \frac{\int_0^{\theta_0} g(\theta) \sum_{f=a+1}^{\infty} h[f, T|\theta] d\theta}{\int_0^{\infty} g(\theta) \sum_{f=a+1}^{\infty} h[f, T|\theta] d\theta} \geq 1 - \alpha^* \quad (6)$$

We let the observed number of failures and time; f and T , respectively, be the observed statistics. Thus, we replace $h(S|\theta)$ with $h[f, T|\theta]$ where the latter may be represented by a discrete distribution such as the Poisson for a practical application.

The summations in the above expressions provide the conditional probability that $0 < f \leq a$ and $f \geq a + 1$, respectively, as required by the left hand sides. It is clear from examining Eqs 5 and 6 that the MTBF, θ is a random variable where $0 < \theta < \infty$ and α^* and β^* are the Bayesian producer's and consumer's risks, respectively. This is a point of departure from the classical approach wherein the risks concern the probability of passing or failing the test itself as implied by Eqs 1 and 2. Note that $h[f, T|\theta]$ may be a continuous density if necessary, requiring only that the summation signs in Eqs 5 and 6 may be replaced by integral signs.

Upon stating the inputs θ_0 , θ_1 , α^* and β^* , simultaneous solution of Eqs 5 and 6 will generate many feasible solutions (T , a). Moreover, the minimum T , say T_m , and associated a may be determined for the given inputs. Thus, the decision rule is to test the product for a total of T_m hours; if a or less failures are observed the product is accepted, otherwise, it is rejected. Note that this decision rule is the same as in the classical case and upon varying the inputs it is possible to develop a table of reliability demonstration plans.

The major differences between the classical and Bayesian approaches are now evident. First, the prior distribution concept does not exist in the classical interpretation. In the Bayesian framework it is of paramount importance. This implies that if efficiencies are to be realized through Bayesian methodology careful attention must be given to the selection of the prior distribution. Second, the interpretation of the consumer's and producer's risks is fundamentally different. In the classical case the consumer wishes to minimize the probability that the reliability demonstration test accepts a bad product. The producer wishes to minimize the probability that the reliability demonstration

rejects a good product. In the classical approach, given the inputs $\gamma_1, \gamma_2, \theta_1, \theta_0$, and the underlying model, all of which generate the decision rule, the relative frequency with which the accept or reject decision is made reduces to a dependence upon the true, but unknown, value of θ .

In the Bayesian framework the consumer wishes to have maximum assurance that the product that has been accepted and he has received will meet his MTBF requirement for his application; while the producer wishes maximum assurance upon product rejection that the MTBF is indeed less than the value θ_0 . The decision rule is generated by the inputs $\alpha^*, \beta^*, \theta_1, \theta_0$, and the conditional sampling distribution as in the classical case, however, thereafter, it depends entirely on the prior distribution employed. It is evident, therefore, that the classical decision rule is oriented toward the value of the MTBF going into the demonstration test. The Bayesian viewpoint is oriented toward having both the producer and the consumer be primarily concerned with the probabilistic assertions they may make *after* the reliability test has been conducted and a decision regarding the product has been reached. The classical view might be meaningful for the producer, for he is, of course, greatly interested in not having a product rejected and, therefore, attempts to enter a reliability test with as high an MTBF as possible within scheduling and budget constraints. However, the classical approach appears to be less meaningful for the consumer since his primary concern is to have high assurance, within scheduling and budget constraints, that the product he receives will meet his requirement.

The third major difference is of an obvious mathematical nature. In the classical case, the true MTBF, θ , is an unknown constant whereas in the Bayesian case, θ is a random variable having customary properties including a probability density function describing it. This permits the different interpretations discussed earlier and again points out the importance of the prior distribution.

THE PRIOR DISTRIBUTION AND PARAMETER SELECTION

Application of Bayes' Theorem to reliability demonstration requires that all objective and subjective MTBF information, be integrated to produce a prior density of the MTBF before the demonstration test is conducted. Subsequently, to bridge the gap between this information and a convenient mathematical form, one of the idealized statistical probability density functions is "fitted" to the information. A number of density functions are candidates for such a fit. Selection of a particular one can be based on statistical tests for goodness of fit or other more subjective considerations. The density function selected to represent the prior density herein is the inverted gamma.

The inverted gamma prior distribution may be written as

$$g(\theta) = \left\{ \alpha^\beta / \left[\Gamma(\beta)(\theta - A)^{\beta+1} \right] \right\} e^{-\alpha/(\theta - A)}, \theta > 0 \quad (7)$$

where θ , the MTBF, is a random variable

$$\theta > 0$$

$$\alpha, \beta > 0$$

$$A \geq 0$$

Note that α and β are the scale and shape parameters, respectively. Figure 1 illustrates some of the family of prior densities that can be represented by the inverted gamma. All densities shown have origins at $A = 0$. If $A > 0$, the shapes remain the same, then each density would be translated to the right by a distance, say A , and it is implied that the MTBF cannot be less than A .

The inverted gamma was selected for the following reasons:

- A. There is good reason to believe that the distribution of the operational, (true) MTBF of electronic systems tends toward positive skewness, [5] and [9]. With this in mind, the inverted gamma is particularly suitable since upon properly selecting its parameters it conveniently fits positively skewed densities. Also, under certain parameter conditions, the inverted gamma approaches symmetry.
- B. The inverted gamma is mathematically tractable within Bayes' Theorem.
- C. Parameter selection for the inverted gamma is relatively simple.

PARAMETER SELECTION

Specifying values for α and β , the scale and shape parameters, respectively, of Eq 8 is tantamount to selecting a particular probability density function from the inverted gamma family. A procedure for specifying α and β is developed by recognizing that if the value of any two fractile points of an inverted gamma density are known, the distribution can be completely specified. If we let the predicted MTBF, θ^* , be the median, i.e., the 50th fractile, $F_{0.50}$, that cuts off the lower 50 percent of the prior distribution it is implied that the predictor of $F_{0.50}$ believes that the probability is one-half that the true MTBF is less than $F_{0.50}$. A second fractile point may be determined by subjectively selecting an MTBF value such that the odds are 3:1 that the true MTBF is less than the value selected. The latter value, $F_{0.75}$, cuts off the lower 75 percent of the distribution. Since the cumulative distribution of the inverse gamma with the origin at zero is

$$G(x; \alpha, \beta) = \int_0^x [1/\Gamma(\beta)] \alpha^\beta e^{-\alpha/\theta} (1/\theta)^{\beta+1} d\theta = \sum_{j=0}^{\beta-1} (1/j!) e^{-\alpha/x} (\alpha/x)^j \quad (8)$$

it follows that substituting $x = F_{0.50}$ and $x = F_{0.75}$ in Eq 8 yields

$$G[F_{0.50}; \alpha, \beta] = 0.50$$

and

$$G[F_{0.75}; \alpha, \beta] = 0.75$$

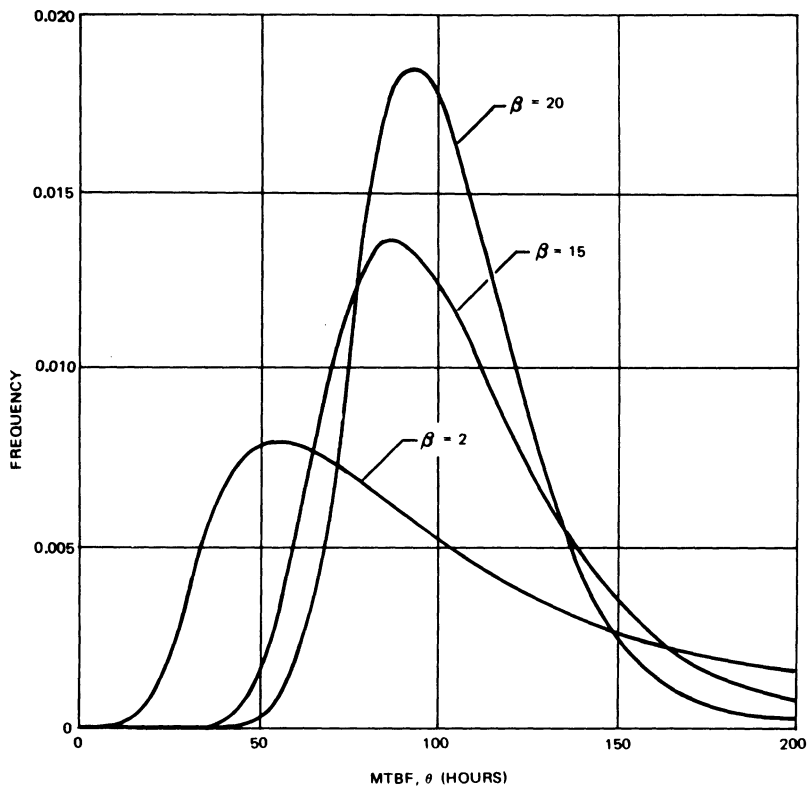


Figure 1. Various Inverted Gamma Densities with Zero Origin and Median at $\theta = 100$

For a selected integer value of β , expression (8) provides unique solutions Z_1 and Z_2 where

$$\begin{aligned} Z_1 &= \alpha/F_{0.50} \\ Z_2 &= \alpha/F_{0.75} \end{aligned} \quad (9)$$

Furthermore, the following ratio will be unique:

$$\begin{aligned} Q &= Z_1/Z_2 \\ &= (\alpha/F_{0.50})/(\alpha/F_{0.75}) \\ &= F_{0.75}/F_{0.50} \end{aligned}$$

Thus, for a selected value of β , a unique value is obtained for Q . If various β values are considered, then corresponding values for Q can be generated. Moreover, if a table of (β_i, Q_i) is constructed, it would be necessary to merely calculate the ratio Q_i to select the corresponding β_i from the table. Once β is determined, α is easily calculated from either expression in Eq 9. Here, the first expression will be used. Since the subscript is no longer necessary,

$$\alpha = ZF_{0.50}. \quad (10)$$

Therefore, the table can be augmented with corresponding values of α_i for given β_i to provide a table of (Q_i, β_i, α_i) . The table that has been constructed (Table 1) is (Q_i, β_i, Z_i) since the proportionality factor Z simplifies the implementation. Table 1, developed with the aid of tables by Pearson [8], presents a range of Q considered useful for practical applications. If the origin is shifted to $A > 0$, all preceding arguments are the same except that $(F_m - A)$ is substituted for all fractile points, F_m , and A is selected subjectively. A detailed parameter selection and evaluation procedure is provided in a subsequent section.

PRIOR DISTRIBUTION EVALUATION

Once a specific density has been selected by virtue of the calculated Q , a useful graph of the cumulative distribution may be generated by substituting β and α where $\alpha = ZF_{0.50}$ into Eq 8, and setting $X = F_{0.50}, F_{0.75}$ and other desired values of F_m , a graph of the probability density function may be constructed. Either or both of these graphs may be examined to judge if the prior density selected does, in fact, represent the prior beliefs of the producer and consumer. Thus, a method of interpreting α and β is available. If the graph is acceptable the inequalities represented by Eqs 5 and 6 may be implemented directly. If the selected prior distribution is not acceptable, new values for $F_{0.50}$ and/or $F_{0.75}$ must be determined which, in turn, will modify Q and, of course, α and β . The new cumulative distribution and the probability density function may then be constructed and examined. Eventually the prior distribution will be acceptable and the process will terminate. It is of interest to note that as Q becomes large, β decreases, causing a more positive skewed distribution. When Q decreases, β increases, forcing the prior toward symmetry. More than two points can, of course be used for the construction of the distribution function. This is discussed in detail in [7, 13, and 14].

Table 1
SELECTED RATIOS OF Q FACILITATING THE ESTIMATION OF β AND
Z FOR THE INVERSE GAMMA PRIOR DISTRIBUTION

Q	β	Z
1.746	2	1.679
1.548	3	2.674
1.449	4	3.672
1.386	5	4.671
1.344	6	5.671
1.312	7	6.671
1.288	8	7.670
1.268	9	8.670
1.252	10	9.670
1.238	11	10.669
1.226	12	11.669
1.215	13	12.669
1.207	14	13.669
1.199	15	14.667

FUNDAMENTAL FORMS

Upon setting the prior distribution equal to the inverted gamma with specific values for α and β and defining the conditional sampling distribution as Poisson, inequalities Eqs 5 and 6 become

$$\frac{\int_A^{\theta_1} \alpha^\beta [1/(\theta-A)]^{\beta+1} \sum_{f=0}^a (1/f!) e^{-(\alpha+T)/(\theta-A)} [T/(\theta-A)]^f d\theta}{\int_A^\infty \alpha^\beta [1/(\theta-A)]^{\beta+1} \sum_{f=0}^a (1/f!) e^{-(\alpha+T)/(\theta-A)} [T/(\theta-A)]^f d\theta} \leq \beta^* \quad (11)$$

$$\frac{\int_A^{\theta_0} \alpha^\beta [1/(\theta-A)]^{\beta+1} \sum_{f=a+1}^\infty (1/f!) e^{-(\alpha+T)/(\theta-A)} [T/(\theta-A)]^f d\theta}{\int_A^\infty \alpha^\beta [1/(\theta-A)]^{\beta+1} \sum_{f=a+1}^\infty (1/f!) e^{-(\alpha+T)/(\theta-A)} [T/(\theta-A)]^f d\theta} \geq 1-\alpha^* \quad (12)$$

If $A = 0$, i.e., the origin is at zero, performing the integration in Eqs 11 and 12 under the appropriate transformations yields

$$\frac{\sum_{f=0}^a (R/S)^f (V!/f!) \sum_{j=0}^V e^{-S/D} (S/D)^j / j!}{\sum_{f=0}^a (R/S)^f (V!/f!)} \leq \beta^* \quad (13)$$

$$\frac{(\beta-1)! \sum_{j=0}^{\beta-1} (1/j!) e^{-W/S} (W/S)^j - (W/S)^\beta \sum_{f=0}^a (R/S)^f (V!/f!) \sum_{j=0}^V (1/j!) e^{-S/D} (S/D)^j}{(\beta-1)! - (W/S)^\beta \sum_{f=0}^a (R/S)^f (V!/f!)} \geq 1-\alpha^* \quad (14)$$

where

$$S = ZP + R$$

$$W = ZP$$

$$V = \beta + f - 1$$

and the following proportionality factors have been defined and inserted for convenience

$$D = \theta_1/\theta_0 \text{ (discrimination ratio)}$$

$$P = \theta^*/\theta_0$$

$$R = T/\theta_0$$

$$Z = \alpha/\theta^*.$$

Thus, for the $A = 0$ case, given α^* , β^* , θ_0 and θ_1 where $\theta_1 < \theta_0$; letting Eq 8 represent the prior density of θ with specified α and β and letting the Poisson distribution represent the conditional sampling distribution, Eqs 13 and 14 may be solved for minimum R and the associated a . Upon determining values for R and a , the following application and decision rule may be invoked: Test the subject system for total of $T = R\theta_0$ hours. If a or less failures are observed, the system is accepted as having fulfilled the pertinent consumer and producer reliability requirements. If the number of failures observed is greater than a , the system is rejected.

By varying the inputs α^* , β^* , θ_0 , θ_1 and selecting specific priors, reliability demonstration test plans may be generated by Eqs 13 and 14. Such plans were constructed for practical ranges of the input variables and are presented in Table 2. If $A > 0$, Eqs 13 and 14 can be modified easily and Table 2 may be employed upon applying a simple adjustment.

IMPLEMENTING THE PLANS

The Bayesian risks α^* and β^* are specified by producer and consumer negotiation as are θ_0 and θ_1 . Routine reliability predictions provide θ^* , thus, $P = \theta^*/\theta_0$ and the discrimination ratio is $D = \theta_1/\theta_0$. In order to select a specific demonstration plan it remains only to select the specific inverted gamma prior density, i.e., α and β . An iterative routine for selecting α and β when the origin is assumed to be zero ($A = 0$) follows:

- A. Let F_m denote the m^{th} fractile point on the prior distribution such that $\Pr(\theta < F_m) = m$. Let $F_{0.50}$ and $F_{0.75}$ be two such points. Set the predicted MTBF, θ^* , equal to $F_{0.50}$, i.e., the median.
- B. Subjectively select an MTBF value such that the odds are 3:1 that the true MTBF is less than the value. Denote the selected value $F_{0.75}$.
- C. Calculate the ratio $Q = F_{0.75}/F_{0.50}$.
- D. Enter Table 1 at the nearest value of the calculated ratio, Q , and read the values for β and Z directly. Note that $\alpha = Z\theta^*$.
- E. Substitute β and α (where $\alpha = Z\theta^*$) into $G[x; \alpha, \beta]$. Set $x = F_{0.50}$, $F_{0.75}$ and other desired values of F_m and graph the resulting cumulative distribution functions (or probability density function if desired). If the resulting graph satisfies the subjective beliefs of the producer and consumer, go to Item F. If the graphed prior density is unsatisfactory select new values for $F_{0.50}$ and/or $F_{0.75}$ and return to Item C.

F. Enter Table 2 with α^* , β^* , β , Z , P and D , and read R and a directly.

The minimum required test time is $T_m = R\theta_0$ and the acceptance number is a failures.

If it is determined beforehand that it is more meaningful that the origin of the inverted gamma prior is at some value $A > 0$, the preceding iterative routine applies with two modifications. First, all references to the F_m must be replaced by $(F_m - A)$ and θ_1 , θ_0 , and θ^* must be replaced by $(\theta_1 - A)$, $(\theta_0 - A)$, and $(\theta^* - A)$, respectively. Second, the origin of the graph of the cumulative distribution function (Item E) must be labeled A .

EXAMPLES

Example 1

Initial conditions: The prior density is represented by the inverted gamma with $A = 0$, $\theta_1 = 700$ hours, $\theta_0 = 1,000$ hours; and $(\alpha^*, \beta^*) = (0.1, 0.1)$.

- A. Let $\theta^* = F_{0.50} = 1,000$ hours.
- B. Let $F_{0.75} = 1,450$ hours.
- C. $Q = 1,450/1,000 = 1.45$.
- D. From Table 1, $(\beta, Z) = (4, 3.672)$.
- E. Let the resulting cumulative prior distribution and prior density function, shown in Figures 2 and 3, respectively, be acceptable to the consumer and producer.
- F. Enter Table 2 with $(\alpha^*, \beta^*, \beta, Z, P, D) = (0.1, 0.1, 4, 3.672, 1.0, 0.7)$ which yields $(R, a) = (5.05, 7)$.

Interpretation: Test the system for $R\theta_0 = 5,050$ hours. If 7 or less failures are observed accept the system, otherwise reject it.

Example 2

Initial conditions: Same as in Example 1 except that $A = 100$ hours, $\theta_1 = 770$ hours, $\theta_0 = 1,060$ hours and $(\alpha^*, \beta^*) = (0.1, 0.1)$.

- A. Let $\theta^* = F_{0.50} - A = 960$ hours.
- B. Let $F_{0.75} - A = 1,300$ hours.
- C. $Q = 1,300/960 = 1.354$. (Note $D = 670/960 = 0.7$)
- D. From Table 1, $(\beta, Z) = (6, 5.671)$.

- E. Let the resulting cumulative prior distribution and prior density function be acceptable to the consumer and producer.
- F. Enter Table 2 with $(\alpha^*, \beta^*, \beta, Z, P, D) = (0.1, 0.1, 6, 5.671, 1.0, 0.7)$ which yields $(R, a) = (4.68, 7)$.

Interpretation: Test the system for $R\theta_0 = 4,961$ hours. If 7 or less failures are observed accept the system, otherwise reject it.

COMPARISONS AND SUMMARY

A typical example of the relatively smaller test time requirement due to the Bayesian approach as opposed to two common classical plans over a range of $P = \theta^*/\theta_0$ for risks of $(\gamma_1, \gamma_2) = (0.1, 0.1)$, $(\alpha^*, \beta) = (0.1, 0.1)$ and a discrimination ratio $D = \theta_1/\theta_0 = 0.8$ is exhibited in Figure 4. Although the plans are not strictly comparable due to the different interpretation of risk, Figure 4 shows the respective test time requirements if the inputs were interpreted in the sense necessary to apply each method. The single sample plan was generated by Eqs 3 and 4 and provides a single test time requirement for all P . The sequential test's *expected test time* is from an AGREE type of plan and here it is necessary to interpret θ^* of $P = \theta^*/\theta_0$ as the classical "true" MTBF. The discrimination ratio for the later case would be quoted as 1.25:1 in AGREE parlance. The savings due to the Bayesian approach are evident from Figure 4. Additional investigations have shown that a decrease (increase) in the Bayesian risk or an increase (decrease) in D will effect an increase (decrease) in the test time requirement. This relationship parallels classical sensitivity and was expected.

It is significant to note that in many cases no testing is required by the Bayesian plans, the product is judged acceptable, and both producer and consumer requirements are satisfied. This result is due to the influence of a "good" prior distribution. That is, the shape of the prior density is such that the requirements are fulfilled without additional information from a demonstration test. This further points out the importance of prior distribution selection. Some significant work in the development of useful prior distributions for the MTBF of electronic hardware has been conducted by Feduccia and Klion [5] and Schafer, et al [9]. Also of interest is the work of Lin and Schick [7], and Winkler [13], [14]. These efforts tend to indicate that the all important prior distribution concept is indeed capable of meaningful, practical interpretation. Thus, the savings in reliability demonstration test time requirements typified in this paper are a very real possibility.

ACKNOWLEDGEMENT

The authors wish to express appreciation to Bruce J. Dimock for his contributions to the development of the computer program which provided Table 2.

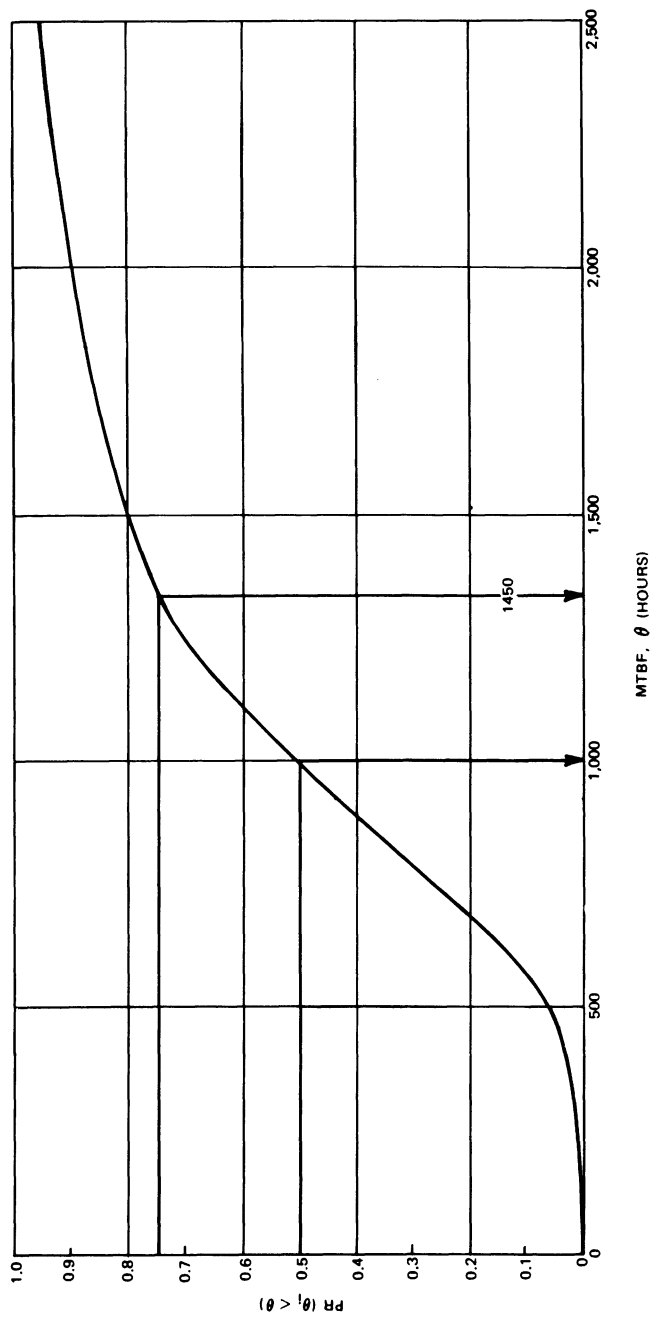


Figure 2. Cumulative Distribution for the Selected Inverted Gamma Prior Distribution ($\beta = \dots$, $Z = 3$, $F_{.50} = 1,000$ Hours and $F_{.75} = 1,450$ Hours)

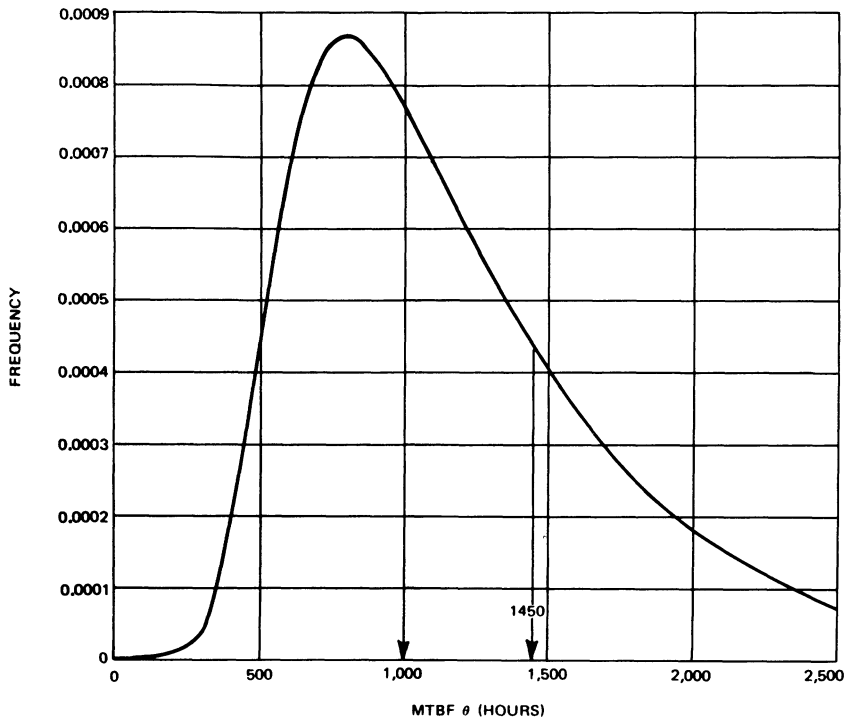


Figure 3. Selected Inverted Gamma Prior Distribution ($\beta = 4$, $Z = 3.7$, $F_{.50} = 1000$ Hours and $F_{.75} = 1450$ Hours)

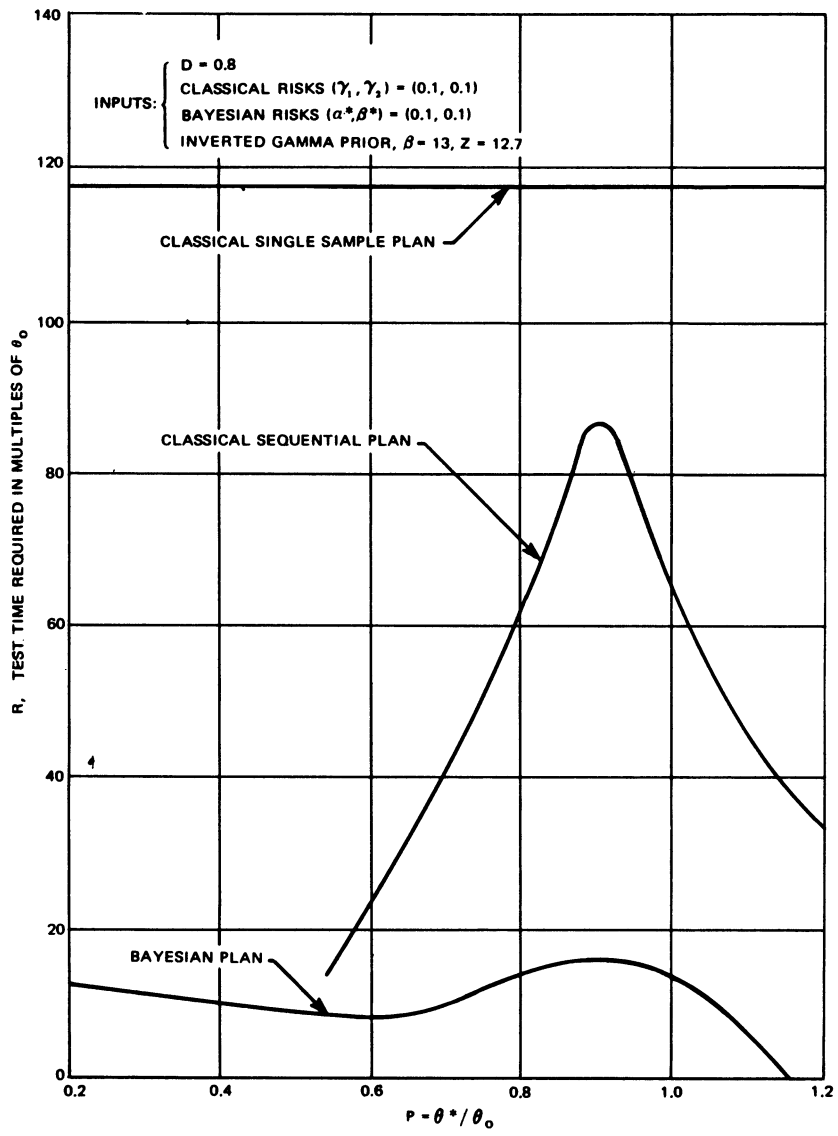


Figure 4. Comparison of Test Time Requirements for Bayesian Plans and Two Classical Plans.

Table 2

RELIABILITY DEMONSTRATION TEST PLANS BASED ON THE
INVERTED GAMMA PRIOR DISTRIBUTION
(R = UPPER TABLE VALUE, a = LOWER TABLE VALUE) (Continued)

$\beta = 6, Z = 5.7 \quad \alpha^* = 0.05, \beta^* = 0.05$					$\beta = 13, Z = 12.7 \quad \alpha^* = 0.05, \beta^* = 0.05$				
D	0.8	0.7	0.5	0.8	D	0.8	0.7	0.5	0.3
P					P				
1.7	†	†	†	†	1.7	†	†	†	†
1.5	†	†	†	†	1.5	†	†	†	†
1.2	22.76 29	8.21 14	†	†	1.2	15.25 25	†	†	†
1.1	25.64 31	10.97 16	†	†	1.1	27.68 36	†	†	†
1.0	26.94 31	12.82 17	†	†	1.0	33.70 40	11.00 17	†	†
0.9	28.05 31	13.86 17	1.31 4	†	0.9	36.08 40	14.83 19	†	†
0.8	26.67 28	14.10 16	2.83 5	†	0.8	34.89 36	16.09 18	†	†
0.7	23.55 23	13.52 14	3.66 5	†	0.7	27.66 25	14.19 13	1.68 2	†
0.5	11.73 7	7.61 4	2.96 1	†	0.5	9.22 0	7.28 0	3.39 0	†

†No Testing Required.

Table 2

RELIABILITY DEMONSTRATION TEST PLANS BASED ON THE
INVERTED GAMMA PRIOR DISTRIBUTION
(R = UPPER TABLE VALUE, a = LOWER TABLE VALUE) (Continued)

$\beta = 3, Z = 2.7 \quad \alpha^* = 0.2, \beta^* = 0.2$					$\beta = 4, Z = 3.7 \quad \alpha^* = 0.2, \beta^* = 0.2$				
D	0.8	0.7	0.5	0.3	D	0.8	0.7	0.5	0.3
P					P				
1.7	†	†	†	†	1.7	†	†	†	†
1.5	†	†	†	†	1.5	†	†	†	†
1.2	0.95 2	†	†	†	1.2	0.22 2	†	†	†
1.1	1.44 2	0.48 2	†	†	1.1	1.74 3	†	†	†
1.0	1.84 2	1.09 2	†	†	1.0	2.39 3	0.84 2	†	†
0.9	2.20 2	1.04 1	†	†	0.9	2.32 2	1.47 2	†	†
0.8	2.54 2	1.37 1	†	†	0.8	2.77 2	1.44 1	†	†
0.7	2.21 1	1.13 0	0.27 0	†	0.7	2.52 1	1.30 0	0.19 0	†
0.5	2.09 0	1.66 0	0.81 0	†	0.5	2.58 0	2.03 0	0.93 0	†

†No Testing Required.

REFERENCES

1. H. Balaban. A Bayesian Approach to Reliability Demonstration. Annals of Reliability and Maintainability Conference, 8th Conference, 1969, P 497-504.
2. A. J. Bonis. Bayesian Reliability Demonstration Plans. Annals of Reliability and Maintainability Conference, 5th Conference, 1966, P 861-873.
3. A. M. Breipohl, R. R. Prairie, and W. J. Zimmer. A Consideration of The Bayesian Approach in Reliability Evaluation. I.E.E.E. Transactions in Reliability, Vol. 14, October 1965, P 107-113.
4. C. H. DeWitt, R. H. Myers, K. L. Wong, and E. G. N. Yem. Reliability Prediction and Demonstration for Airborne Electronics. RADC Contract Number F30602-67-C-0221, April, 1968.
5. A. J. Feduccia and J. Klion. How Accurate are Reliability Predictions? Proceedings, Annual Symposium on Reliability, 1968, P 280-297.
6. C. W. Hamilton. Bayesian Procedures and Reliability Information. Second Annual Aerospace Reliability and Maintainability Conference, 1963, P 278-283.
7. C. Y. Lin and G. J. Schick. On-Line Console-Aided Assessment of Prior Distributions. Ninth Reliability and Maintainability Conference, Detroit, Michigan, 22-20 July 1970.
8. Karl Pearson. Tables of the Incomplete Gamma Function. University Press, Cambridge, England, 1946.
9. R. E. Schaefer, J. Collins, M. L. Luden, et al. Bayesian Reliability Demonstrations: Phase 1 – Data for the A Priori Distribution. RADC-TR-69-389, Final Technical Report, February, 1970.
10. R. E. Schaefer and N. D. Singpurwalla. A Sequential Bayes Procedure for Reliability Demonstration. Annals of Reliability and Maintainability Conference, 8th Conference, 1969, P 507-514.
11. Reliability Tests: Exponential Distribution. U. S. Department of Defense, MIL-STD-781B, Washington, D. C., Government Printing Office, 1967.
12. A. Wald. Sequential Analysis. John Wiley and Sons, Inc., New York, 1947.
13. R. L. Winkler. The Assessment of Prior Distributions in Bayesian Analysis. *Journal of the American Statistical Association*, Vol. 62, No. 319, September 1967, P 776-800.
14. R. L. Winkler. The Qualification of Judgement: Some Methodological Studies. *Journal of the American Statistical Association*, Vol. 62, No. 320, December, 1967, P 1105-1120.

Lösungsvorschläge für stochastische Zielprogramme von W. Rödder, Aachen

Übersicht: Für ein stochastisches Zielprogramm der Form

$$\begin{aligned} Gx + \bar{V} &\geq y^* \\ \text{s. d. } Ax &\leq b \\ x &\geq 0 \end{aligned}$$

werden deterministische Ersatzprobleme vorgeschlagen, indem

- 1.) die Wahrscheinlichkeit der Zielerreichung maximiert bzw.
- 2.) die Wahrscheinlichkeit der Nicht-Zielerreichung minimiert wird.

Es werden Familien von Verteilungen für \bar{V} angegeben, bei denen diese Aufgaben zu konvexen bzw. konkaven Optimierungsproblemen führen. Ferner lassen sich einfache hinreichende Bedingungen für die Existenz von optimalen Lösungen aufstellen.

Einführung in die deterministische Zielprogrammierung

Untersuchungen über die Kriterien, an denen Entscheidungsträger in Unternehmen ihre Entscheidungen messen, führten in den letzten Jahren zu der Hypothese, daß ein Modell, in dem die Entscheidungen mittels einer eindimensionalen "Zielgröße" ge-

ordnet werden (Gewinn, Erlös u.a.), nicht in jedem Fall dem realen Sachverhalt gerecht wird.

Vielmehr zeigt z.B. Heinen (HE), daß der Unternehmer sich simultan von mehreren Zielen leiten läßt. In einem mathematischen Modell spiegelt sich diese Tatsache in einer vektorwertigen Zielfunktion (F) wider.

Sie gibt den funktionellen Zusammenhang zwischen Entscheidungen x und Zielgrößen $F(x)$ an.

Wir wollen i.f. unterstellen, daß sich der Entscheidungsträger hinsichtlich der einzelnen Zielkomponenten

$F_i(x) : i = 1, \dots, r$ Anspruchsniveaus gesetzt hat;

diese Anspruchsniveaus seien bezeichnet mit

$(y_1^*, \dots, y_r^*) =: y^*$.

Sind die möglichen Entscheidungen gewissen Restriktionen

$g_k(x) \leq 0 : k = 1, \dots, m$ unterworfen, läßt sich der

Imperativ

"Triff Deine zulässige Entscheidung so, daß das Anspruchsniveau y^* möglichst gut erreicht wird" formal schreiben als:

$F(x) \geq y^*$

s.d. $g_k(x) \leq 0 \quad k = 1, \dots, m \quad (1)$

Der Ausdruck $F(x) \geq y^*$ kann konkretisiert werden in eine Abweichungsfunktion oder allgemeiner eine Präferenzordnung, mittels der der Entscheidende die Zielerreichungsgrade $F(x)$ bzgl. der Anspruchsniveaus y^* ordnet.

Hierzu finden sich einige Vorschläge bei Charnes und Cooper [CC] S. 215 ff, Ijiri [IJ] S. 55 und Rödter [RO] S. 50 ff.

Unterstellt man in (1), daß $F(x)$ und $g_k(x)$ linear sind, erhält man ein lineares Zielprogramm der Form

$$Gx \geq y^*$$

$$\text{s.d. } Ax \leq b$$

$$x \geq 0$$

(2)

(Hierbei sei: $x \in \mathbb{R}^n$, $y \in \mathbb{R}^r$, $b \in \mathbb{R}^m$, G eine $r \times n$ - Matrix und A eine $m \times n$ - Matrix).

Als einfaches Beispiel für (2) mag folgender Sachverhalt dienen (vgl. [IJ] S. 45) :

In einer Firma arbeiten 2 Gruppen mit verschiedenen Produktionsraten;

Gruppe 1 stelle pro Stunde 1 Einheit

Gruppe 2 " " " 0,6 Einheiten her.

Pro Tag sollen insgesamt - falls möglich - mindestens 12 Einheiten hergestellt werden, und jede Gruppe sollte möglichst höchstens 8 Stunden arbeiten.

In das Modell (2) gefaßt ergibt das:

$$\begin{pmatrix} 1 & 0,6 \\ -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \geq \begin{pmatrix} 12 \\ -8 \\ -8 \end{pmatrix}$$

$$\text{s.d. } x_i \geq 0 \quad i = 1, 2$$

Stochastische Lineare Zielprogrammierung

In der Literatur zur stochastischen linearen Programmierung wurde hinreichend diskutiert, daß in einem Modell der Form

$$\begin{aligned} \text{Max} \quad & cx \\ \text{s.d.} \quad & Ax \leq b \\ & x \geq 0 \end{aligned} \quad (3)$$

die Komponenten des Vektors (c, A, b) gewöhnlich nur Schätzungen der sich in der Zukunft realisierenden Koeffizienten sind. Liegen die Informationen über diese Koeffizienten in Form von Wahrscheinlichkeitsverteilungen vor, beschreibt ein stochastisches L.P.-Modell

$$\begin{aligned} \text{Max} \quad & \bar{c}x \\ \text{s.d.} \quad & \bar{A}x \leq \bar{b} \\ & x \geq 0 \end{aligned}$$

(wobei $\bar{c}, \bar{A}, \bar{b}$) ein Zufallsvektor ist) besser den realen Sachverhalt als (3).

Diese Überlegungen verlaufen völlig analog für lineare Zielprogramme, die bei stochastischem $(\bar{c}, \bar{A}, \bar{b})$ die folgende Form haben:

$$\begin{aligned} \bar{c}x &\geq y^* \\ \text{s.d.} \quad & \bar{A}x \leq \bar{b} \\ & x \geq 0 \end{aligned} \quad (4)$$

Einschränkend soll i.f. angenommen werden, daß (4) für den Fall zu lösen ist, daß die Restriktionen deterministisch sind und die Zielformulierung

$Gx + \bar{V} \geq y^*$ ist, wobei G fest und \bar{V} ein r -dimensionaler Zufallsvektor ist.

Insgesamt ergibt sich also das Problem

$$Gx + \bar{V} \geq y^*$$

$$\text{s.d. } Ax = b \quad (4a)$$

$$x \geq 0$$

1. Maximierung der Wahrscheinlichkeit der Zielerreichung

Die Aufgabe (4a) bedarf einer Interpretation, da der Imperativ "Wähle eine zulässige Entscheidung so, daß der Zufallsvektor $Gx + \bar{V}$ möglichst $\geq y^*$ ist" zunächst einmal sinnlos ist.

Wir nehmen i.f. an, der Entscheidungsträger habe eine konkrete Nutzenkonzeption, und zwar ordne er solchen Ergebnissen von

Entscheidung x und

Realisation der "Umwelt" v^0

den Nutzen 1 zu, falls $Gx + v^0 \geq y^*$

und 0 sonst.

Die Maximierung des erwarteten Nutzens ergibt dann das deterministische Entscheidungsproblem

$$\text{Max } P \{ Gx + \bar{V} \geq y^* \}$$

$$\text{s.d. } Ax \leq b$$

$$x \geq 0$$

(5)

Eine solche Nutzenkonzeption ist z.B. sinnvoll, wenn eine Nicht-Zielerreichung in auch nur einer Komponenten existenzgefährdend für den Entscheidungsträger ist.

Bei unternehmerischen Entscheidungen könnten folgende Beispiele genannt werden: Die y_i^* sind Liquiditätsforderungen oder sie sind in einem Produktionsmodell die Auftragsmengen eines stark dominierenden Kunden, bei deren Nichterreichen ein Abbruch der wirtschaftlichen Beziehungen droht!

Sind $F_i(w)$ die Verteilungsfunktionen der stetig verteilten, unabhängigen Zufallsvariablen \bar{V}_i , gilt:

$$P \{Gx + \bar{V} \geq y^*\} = \prod_{i=1}^r P \{(Gx + \bar{V})_i \geq y_i^*\} =$$

$$\prod_{i=1}^r P \{ \bar{V}_i \geq y_i^* - (Gx)_i \} = \prod_{i=1}^r (1 - F_i(y_i^* - (Gx)_i))$$

Unter den genannten Voraussetzungen ergibt sich für (5) also mit $z_i := y_i^* - (Gx)_i$ und nach Logarithmierung der Zielfunktion:

$$\text{Max} \quad \sum_{i=1}^r \ln (1 - F_i(z_i))$$

$$\text{s.d.} \quad x_i = y_i^* - (Gx)_i \quad i \in \{1, \dots, r\}$$

$$Ax \leq b$$

(6)

$$x \geq 0$$

Wir geben i.f. hinreichende Bedingungen für die Konkavität der Zielfunktion und die Existenz einer optimalen Lösung von (6) an.

Satz 1: Sind die $F_i(z_i)$ stetig differenzierbar und sind $F_i'(z_i) = f_i(z_i)$ Polya - Typ 2 Dichten, ist (6) eine konkave Maximierungsaufgabe.

$$\text{Bew. : } \frac{d}{dz_i} \ln(1 - F_i(z_i)) = - \frac{f_i(z_i)}{1 - F_i(z_i)}$$

Nun gilt nach [BMP], S. 378:

sind $f_i(z_i)$ Polya-Typ 2 Dichten, so ist

$$\frac{f_i(z_i)}{1 - F_i(z_i)} \quad \text{monoton nichtfallend; daraus folgt die Behauptung}$$

Für eine weite Klasse von Verteilungen sind die Bedingungen dieses Satzes erfüllt. So haben z.B. folgende Verteilungen Polya-Typ 2 Dichten.

Die Gammaverteilung, die Weibullverteilung (vgl. [BMP]), S. 378) mit den Dichten.

$$\lambda (\lambda t)^{\alpha-1} \exp(-\lambda t) / \Gamma(\alpha) \quad (\alpha > 0, t \geq 0)$$

$$\lambda \alpha t^{\alpha-1} \exp(-\lambda t^\alpha) \quad (\lambda > 0, t \geq 0)$$

für $\alpha \geq 1$.

die Normalverteilung, die F-Verteilung und die t-Verteilung (vgl. [KAR], S. 124 f).

Wir wenden uns jetzt der Existenz einer optimalen Lösung von (6) zu: Da die Zielfunktion nach oben beschränkt ist,

muß nur garantiert werden, daß es überhaupt ein zulässiges x gibt, für das alle $\ln(1 - F_i(z_i)) > -\infty$ sind. Genauer gilt folgender

Satz 2: Ist $K := \{x : Ax \leq b, x \geq 0\}$ kompakt und $F_i(z_i)$ stetig für $i \in \{1, \dots, r\}$, so ist hinreichend für die Existenz einer optimalen Lösung von (6), daß es ein $x \in K$ gibt, so daß für jedes $i \in \{1, \dots, r\}$ die Wahrscheinlichkeit, das Ziel y_i^* zu erreichen, positiv ist.

Bew.: Da $\ln(1 - F_i(z_i)) \leq 0$ (also nach oben beschränkt), ist hinreichend für die Existenz einer optimalen Lösung von (6), daß

$$1 - F_i(z_i) > 0 \quad i \in \{1, \dots, r\} \quad \text{bzw.}$$

$$\prod_{i=1}^r (1 - F_i(z_i)) > 0$$

Hat z.B. \bar{V} auf den ganzen R^r eine positive Dichte, ist die Bedingung des Satzes erfüllt.

Beispiel einer speziellen Verteilung der \bar{V}_i :

Ist \bar{V}_i auf $[-\frac{1}{\lambda_i}, \infty)$ ($\lambda_i > 0$) $i \in \{1, \dots, r\}$

exponentiell verteilt, ergibt sich für (6) folgendes Optimierungsproblem:

$$\text{-Min} \quad \sum_{i=1} y_i$$

$$\text{s.d.} \quad y_i = \left(1 + \lambda_i (y_i^* - (Gx)_i) \right)^+ \quad 1$$

$$Ax \leq b$$

$$x \geq 0$$

Dies ist ein lineares deterministisches Optimierungsproblem, das z.B. mit dem Simplexalgorithmus gelöst werden kann.

2. Minimierung der Ziel-Nichterreichung

Die Aufgabe (4a) kann eine andere Interpretation als unter 1. erfahren.

So unterstellen wir dem Entscheidungsträger jetzt folgende Nutzenkonzeption:

Einer Entscheidung x und der Realisation der "Umwelt" G^0 ordne der Entscheidungsträger den Nutzen 0 zu, falls $G^0 x < y^*$, und sonst den Nutzen 1.

Das besagt, der Entscheidende ist zufrieden, sobald auch nur in einer Komponenten das Ziel erreicht wird. Diese Bewertung der Zielvariablen ist dann sinnvoll, wenn die Ziele miteinander völlig substituierbar sind.

Bei den gleichen Voraussetzungen und Bezeichnungen wie unter 1 ergibt sich folgendes Minimierungsproblem:

$$1 \quad \text{Wobei für } l \in \mathbb{R} \quad l^+ = \max(0, l)$$

$$\begin{aligned}
 & r \\
 \text{Min } & \sum_{i=1} \ln F_i(z_i) \\
 \text{s.d. } & z_i = y_i^* - (Gx)_i \\
 & Ax \leq b \\
 & x = 0
 \end{aligned} \tag{7}$$

Analog zu 1. läßt sich wieder eine hinreichende Bedingung für die Existenz einer optimalen Lösung zu (7) angeben:

Satz 3: Sind die $F_i(z_i)$ stetig, K^1 kompakt, so ist hinreichend für die Existenz einer optimalen Lösung von (7), daß es kein $x \in K$ gibt, so daß das Ziel y_i^* für irgendein i für jede Realisation von \bar{V}_i erreicht wird.

(Der Beweis erübrigt sich).

Das Problem (7) ist ein konkaves Minimierungsproblem, falls die \bar{V}_i $i \in \{1, \dots, r\}$ gamma-, weibull- oder normalverteilt sind. Bei den ersten beiden Verteilungen folgt diese Tatsache aus $f'(z_i) \leq 0$ ($\Rightarrow \frac{d^2}{dz_i^2} \ln F_i(z_i) \leq 0$), bei der

Normalverteilung zeigt man die Behauptung mittels einer Diskussion der 2. Ableitung

1 siehe Satz 2

Literaturverzeichnis

- [BMP] Barlow, R.E., Properties of probability distributions with monotone hazard rate
Marshall, A.W., Ann. of Math. Stat.. Vol. 34 (1961)
Proshan, F. Nr. 2, S. 375 ff.
- [CC] Charmes, A., : Management models and industrial
Cooper, W.W. applications of linear Programming
Vol. 1, New York, London, Sydney
1961
- [HE] Heinen, E. : Das Zielsystem der Unternehmung,
Wiesbaden 1966
- [IJ] Ijiri, Y. : Management goals and accounting
for control,
Chicago 1964
- [KAR] Karlin, S. : Decision theory for polya type distributions,
Case of two actions I,
Third Berkeley Symposium on Probability and Statistics,
Vol. I (1956), S. 115 ff.
- [RO] Rödder, W. : Deterministisches und stochastisches
Zielprogrammieren,
Dissertation, Aachen 1971

Zusammenfassung

Die Arbeit verfolgt zwei Ziele. Zunächst wird unter Verwendung des Dynamischen Programmierens eine neue Herleitung des Theil'schen Satzes über dynamische Sicherheitsäquivalente gegeben. Sodann wird ein Satz hergeleitet, der die Existenz dynamischer Sicherheitsäquivalente auch für nichtquadratische Kriterien nachweist.

1. Einleitung

Der Begriff dynamisches Sicherheitsäquivalent wurde bereits 1957 von H. Theil [7] geprägt. Grobgesprochen besagt er, daß in gewissen stochastisch dynamischen Optimierungsproblemen die auftretenden stochastischen Prozesse durch bestimmte deterministische Größen, die sogenannten dynamischen Sicherheitsäquivalente, ersetzt werden können. Oder präziser, Theil konnte den folgenden Satz beweisen [8]: "Die optimale Politik eines durch eine lineare Zustandstransformationsgleichung und ein quadratisches Kriterium beschriebenen (dynamisch) stochastischen Optimierungsproblems ändert sich nicht, wenn man die auftretenden Zufallsvariablen durch gewisse bedingte Erwartungswerte ersetzt." Diese bedingten Erwartungswerte nennt man dynamische Sicherheitsäquivalente.

Wir wollen uns diesen Satz etwas veranschaulichen. Gegeben sei das Rückkopplungs-Regelungsproblem der Fig. 1.1

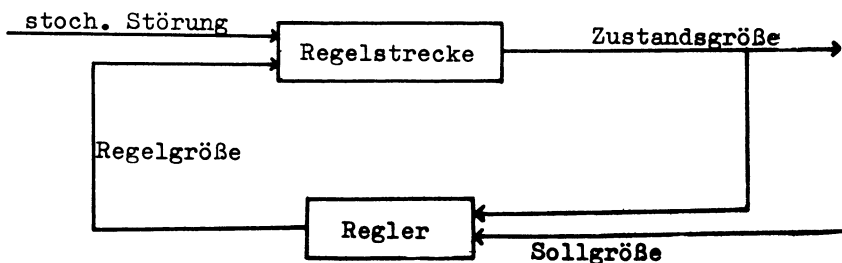


Fig. 1.1

Dann besagt der Theil'sche Satz folgendes: Für den Fall, daß die Regelstrecke eine lineare Transformation beschreibt und das Kostenkriterium quadratisch in der Zustands- und Regelgröße ist, kann die stochastische Störung des Systems durch dynamische Sicherheitsäquivalente ersetzt werden, d.h. durch bedingte Erwartungswerte der Störvariablen unter der Bedingung der bis zum gegenwärtigen Zeitpunkt erhältlichen Information.

Noch deutlicher wird die Bedeutung des Theil'schen Theorems für den Unternehmensforscher, wenn man den obigen Regelungsvorgang als einen Produktionsglättungsvorgang interpretiert. In diesem Falle spezialisiert sich Fig. 1.1 zu Fig. 1.2

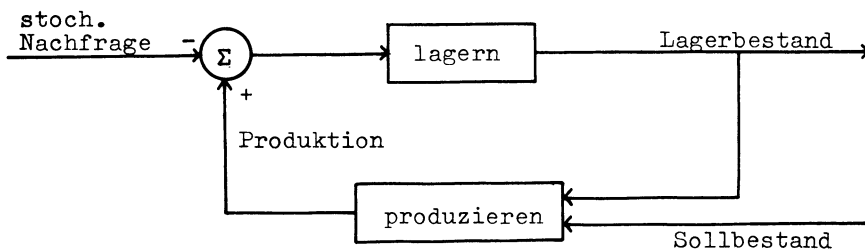


Fig. 1.2

Hier besagt der Satz von Theil, daß unter der per definitionem (backlog-Fall) gegebenen linearen Bilanzgleichung und bei Vorliegen eines quadratischen Kostenzusammenhangs der stochastische Nachfrageprozeß **für die Optimierung durch eine Folge bedingter Erwartungen ersetzt werden kann. Diese bedingten Erwartungswerte sind aber nichts anderes als Erwartungsprognosen der Nachfrage.**

Damit wird aber die Bedeutung des Nachweises der Existenz dynamischer Sicherheitsäquivalente (zumindest innerhalb der Lagerhaltungstheorie) klar ersichtlich. Zeigt doch der Satz von Theil, daß bisher nur sichergestellt ist, daß das in weiten Bereichen der Lagerhaltungstheorie meist unreflektiert verwandte Verfahren der Nachfrageprognostizierung lediglich **im Falle quadratischer Kostenabhängigkeiten nicht suboptimal ist. Quadratische Kriterien treten aber in Problemen der**

Unternehmensforschung nicht allzu häufig auf. Es wäre daher von größtem Interesse, den Theil'schen Satz in der Weise zu erweitern, daß auch nichtquadratische Kostenabhängigkeiten auf dynamische Sicherheitsäquivalente führen.

Wir wollen in dieser Arbeit zwei Probleme behandeln. In einem ersten Teil soll zunächst eine neue Ableitung des Satzes von Theil mit Hilfe des Dynamischen Programmierens gegeben werden. Theil selbst hat zur Ableitung seines Satzes ein quadratisches Variationsverfahren benutzt, das die Inversion einer i.a. hochdimensionalen Matrix erforderlich macht. Es ist daher nicht uninteressant, diesen wichtigen Satz nochmals unter Verwendung des Dynamischen Programmierens abzuleiten. Der zweite Teil der Arbeit ist dann einer Erweiterung des Satzes auf nichtquadratische Kriterien gewidmet.

2. Ein neuer Beweis des Satzes von Theil über die Existenz dynamischer Sicherheitsäquivalente.

A) Das Modell

Wir betrachten das durch die 3 Bestimmungsgrößen a) System, b) Störung und c) Kostenkriterium festgelegte stochastisch dynamische Optimierungsproblem.

a) System

Es seien

x_k : Zustand z.Z. k ($k = 1, \dots, N$)

x_0 : reellwertig vorgegebener Anfangszustand

u_k : Steuerung z.Z. k ($k = 0, \dots, N-1$)

r_k : Störung z.Z. k ($k = 1, \dots, N$)

Diese Variablen seien verknüpft durch die lineare Zustands-transformationsgleichung

$$x_k = x_{k-1} + u_{k-1} + r_k \quad (2.1)$$

Ferner hänge die Steuerung von sämtlichen realisierten (der Messung zugänglichen) Zuständen ab

$$u_{k-1} = \Psi_{k-1}(x_{k-1}, \dots, x_1, x_0) \quad (2.2)$$

Die Gln. (2.1) und (2.2) bilden das "System". Es sei nochmals zum besseren Verständnis des durch (2.2) beschriebenen Rückkopplungsvorganges und insbesondere im Hinblick auf die

Darstellungen der Fig. 1.1 und 1.2 als Blockschaltbild in Fig. 2.1 wiedergegeben

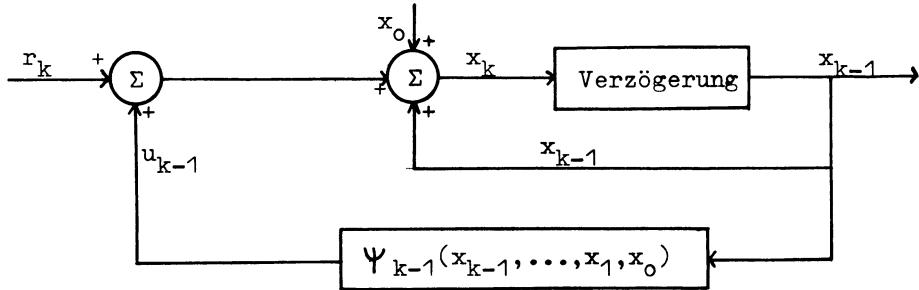


Fig. 2.1

b) Störprozeß

Der Störprozeß $\{r_k\}$ ($k = 1, \dots, N$) sei eine extern vorgegebene Folge von Zufallsvariablen, für die die bedingten Erwartungswerte

$$E \{r_k | r_j, r_{j-1}, \dots, r_{j-i}, x_0\} \quad (k \geq j \geq i \geq 0)$$

existieren mögen.

$$(r_0 = 0)$$

c) Kostenkriterium

Das Kostenkriterium sei gegeben durch den quadratischen Ausdruck

$$K = \sum_{k=1}^N (x_k^2 + \rho^2 u_{k-1}^2) \alpha^{k-1} \quad (2.3)$$

Hierbei bedeutet ρ^2 eine positive Konstante und α einen Diskontfaktor.

Das Optimierungsproblem kann nun folgendermaßen formuliert werden: Es ist eine Folge u_0^0, \dots, u_{N-1}^0 derart zu finden, daß der Erwartungswert der Kosten minimiert werde. Unsere Aufgabe wird nun sein, zu zeigen, daß die Folge $\{u_k^0\}$ nicht von dem Störprozeß $\{r_k\}$, sondern lediglich von gewissen bedingten Erwartungswerten der r_k abhängt. Zunächst noch eine Bemerkung.

Bemerkung

Das soeben definierte Modell ist auch unter den linear-quadratischen Modellen von sehr einfacher Struktur. Es ist aber einsichtig, daß das im folgenden benutzte Verfahren der

Dynamischen Programmierung ohne Schwierigkeiten auch auf multivariate Modelle mit etwas reichhaltigerer linearer (stabiler) Zustandstransformation und komplizierterer quadratischer Kostenfunktion angewandt werden kann.*)

Daß wir gerade dieses Modell gewählt haben, liegt an seiner einfachen Interpretierbarkeit als Produktionsglättungsmodell. Für ein solches Modell wäre x_k zu identifizieren mit dem (positiven bzw. negativen) Lagerbestand am Ende der Periode $(k-1, k]$, r_k stellt (nach Multiplikation mit -1) die Nachfrage während dieser Periode dar (kumuliert in k), und u_{k-1} beschreibt den Produktionsentscheid am Ende der Periode $(k-2, k-1]$. Die Zustandstransformationsgleichung (2.1) ist die bekannte Bilanzgleichung der Lagerhaltungstheorie, und die Kostenfunktion (2.3) beschreibt diskontierte quadratische Lager- und Produktions(abweichungs)kosten während der N Entscheidungsperioden.

B) Lösung des Optimierungsproblems

Zur Lösung des Optimierungsproblems mit Hilfe der Dynamischen Programmierung geht man aus von dem folgenden System von Funktionalgleichungen

$$f_{k+1}(x_k) = \min_{u_k} E\{x_{k+1}^2 + \rho^2 u_k^2 + \alpha f_{k+2}(x_{k+1}) | r_k, r_{k-1}, \dots, r_1, x_0\}$$

($k = 0, 1, \dots, N-2$)

mit der Anfangsgleichung ($k = N-1$)

$$f_N(x_{N-1}) = \min_{u_{N-1}} E\{x_N^2 + \rho^2 u_{N-1}^2 | r_{N-1}, \dots, r_1, x_0\}$$

(Bemerkung: Zur Vereinfachung der Notierung haben wir in der Wertfunktion f_{k+1} die Abhängigkeit von der "Vorgeschichte" $\{r_k, r_{k-1}, \dots, r_1, x_0\}$ nicht gekennzeichnet.)

Damit reduziert sich das obige Optimierungsproblem auf die Lösung der Funktionalgleichungen. Diese Lösung existiert und ist eindeutig. Man erhält nach einer Reihe von Rechenschritten, die im Prinzip in [4] zu finden sind, den folgenden Lösungsalgorithmus:

*) Eine solche Erweiterung wurde als Teil einer Dissertation am Lehrstuhl für Operations Research der FU Berlin bereits durchgeführt.

a) Optimale Politik

$$u_{N-k}^0 = S_{N-k}(x_{N-k} + y_{N-k}) \quad ; \quad (k = 1, \dots, N) \quad (2.4)$$

$$S_{N-k} := \frac{1 - \vartheta^2 \alpha S_{N-k+1}}{\vartheta^2 (\alpha S_{N-k+1} - 1)} \quad ; \quad (k = 2, \dots, N) \quad (2.5)$$

$$Q_{N-k} := \hat{r}(N-k+1|N-k, \dots) + \frac{\hat{T}_{N-k}}{2W_{N-k+1}} \quad ; \quad (k = 2, \dots, N) \quad (2.6)$$

$$\hat{T}_{N-k} := E \{ T_{N-k+1} | r_{N-k}, \dots, r_1, x_0 \} \quad (2.7)$$

$$T_{N-k+1} := -2\vartheta^2 \alpha S_{N-k+1} Q_{N-k+1} \quad (2.8)$$

$$W_{N-k+1} := 1 - \vartheta^2 \alpha S_{N-k+1} \quad (2.9)$$

wo

$$\hat{r}(N-k+1|N-k, \dots) := E \{ r_{N-k+1} | r_{N-k}, \dots, r_1, x_0 \} \quad (2.10)$$

und speziell

$$\hat{r}(k|0, \dots) := E \{ r_k | x_0 \} \quad (2.11)$$

Die Anfangswerte des Algorithmus sind gegeben durch:

$$S_{N-1} := -\frac{1}{1 + \vartheta^2} \quad (2.12)$$

$$Q_{N-1} := \hat{r}(N|N-1, \dots) \quad (2.13)$$

b) Wertfunktionen $f_{N-k+1}(x_{N-k})$

$$\alpha f_{N-k+1}(x_{N-k}) = -\vartheta^2 \alpha S_{N-k} x_{N-k}^2 + T_{N-k} x_{N-k} + R_{N-k} \quad (2.14)$$

$$R_{N-k} := \alpha \left[w_{N-k+1} S_{N-k}^2 + w_{N-k+1} \sigma_{rr}^{(N-k+1|N-k, \dots)} + \hat{T}_r^{(N-k+1)} + \hat{R}_{N-k} \right] \quad (2.15)$$

$$\hat{R}_{N-k} := E \{ R_{N-k+1} | r_{N-k}, \dots, r_1, x_0 \} \quad (2.16)$$

$$\hat{T}_r^{(N-k+1)} := E \{ T_{N-k+1} r_{N-k+1} | r_{N-k}, \dots, r_1, x_0 \} \quad (2.17)$$

$$\sigma_{rr}^{(N-k+1|N-k, \dots)} := E \{ r_{N-k+1}^2 | r_{N-k}, \dots, r_1, x_0 \} \quad (2.18)$$

mit dem Anfangswert

$$R_{N-1} := S_{N-1} \hat{r}^2(N|N-1, \dots) + \sigma_{rr}^{(N|N-1, \dots)} \quad (2.19)$$

Betrachten wir zunächst den Algorithmus für die optimale Politik. Setzt man sukzessive die Gln. (2.5) ... (2.13) in Gl. (2.4) ein, so zeigt sich, daß die optimalen Entscheidungen u_{N-k}^0 lediglich von bedingten Erwartungswerten der Form (2.10) abhängen. Das heißt aber, daß es für die Bestimmung der optimalen Politik $\{u_k^0\}$ lediglich auf die durch (2.10) definierten dynamischen Sicherheitsäquivalente ankommt und nicht auf die volle stochastische Struktur des Störprozesses. M.a.W. man bekäme dieselbe optimale Politik, wenn man von vornherein den stochastischen Störprozeß durch seine dynamischen Sicherheitsäquivalente ersetzt hätte. Damit ist der Theil'sche Satz bewiesen.

3. Ein Satz über die Existenz dynamischer Sicherheitsäquivalente bei nichtquadratischen Kriterien

Wir wenden uns nun der Erweiterung des Theil'schen Satzes auf nichtquadratische Kriterien zu. In jüngster Zeit wurden in verschiedenen Richtungen Anstrengungen unternommen, zu weiteren Aussagen über dynamische Sicherheitsäquivalente zu gelangen. Zu erwähnen sind etwa die Arbeiten von Ziemba [9], Midler [3] und Malinvaud [2]. Ziemba und Midler weisen die Existenz dynamischer Sicherheitsäquivalente nach, falls die Kostenfunktion linear vom Zustand abhängt und weitere weniger

einschneidende Voraussetzungen erfüllt sind. Die lineare Abhängigkeit von den Zustandsvariablen (bei nichtbeschränktem Zustandsraum) stellt aber für die meisten Probleme der Unternehmensforschung eine nicht vertretbare Einschränkung dar. Malinvaud dagegen führt für ein relativ allgemeines nicht-lineares und nichtquadratisches Modell eine Sensitivitätsanalyse durch. Er untersucht die Frage, wie die optimale Politik eines deterministischen Modells sich ändert, wenn gewisse Parameter des Modells in definierter Weise stochastisch werden. Diese Analyse setzt allerdings u.a. eine zweifache Differenzierbarkeit der Kostenfunktion nach den Zustandsvariablen und das Fehlen statischer Nebenbedingungen voraus, beides Voraussetzungen, die in den Anwendungen der Unternehmensforschung nur selten gegeben sein dürften.

Wir wollen nun einen weiteren Satz über dynamische Sicherheitsäquivalente ableiten. Auch dieser Satz enthält noch eine stark einschränkende Voraussetzung, ist aber zumindest den Problemen der Lagerhaltungstheorie, auf die wir unser besonderes Augenmerk richten, besser angepaßt als die zuvor erwähnten Untersuchungen. Drei wesentliche Voraussetzungen bzgl. der Modellkonstituenten a), b) und c) werden wir fordern müssen

V1: Das System sei linear.

V2: Der Störprozeß sei ein Gaußprozeß.

V3: Das Modell befinde sich in einem "eingeschungenen Zustand" (steady state) und lasse ein Kriterium zu, das mittlere Kosten pro Periode beschreibe.

Voraussetzung V1 ist zweifellos am einschneidendsten. Sie besagt, daß nicht nur (wie im Satz von Theil) die Zustandstransformationsgleichung (2.1) als linear vorausgesetzt wird, sondern auch die Funktion Ψ_k in Gl. (2.2). Wie wir wissen, ist z.B. für Lagerhaltungsprobleme vom backlog Typ die Zustandstransformationsgleichung ex definitione linear; kritisch ist daher lediglich die geforderte lineare Abhängigkeit der Politik von den Zustandsvariablen. Wir werden darauf zum Schluß noch einmal zu sprechen kommen.

Die Voraussetzung V2 ist in vielen Fällen nicht kritisch. Wie Box und Jenkins [1] nachgewiesen haben, lassen sich viele ökonomische Zeitreihen als Realisierungen (i.a. höher korrelierter und häufig stationärer) Gaußprozesse interpretieren.

V3 schließlich besagt, daß sich das Kostenkriterium als Erwartungswert einer vom Zustand und der Steuerung abhängigen Funktion $f(x,u)$ darstellen läßt

$$F := E \{ f(x,u) \} \quad (3.1)$$

wobei F also nicht mehr von der Periodennummer abhängt. Ein solcher "eingeschwungener Zustand" wird in vielen Fällen schon nach wenigen Entscheidungsperioden erreicht. (Im Falle einer quadratischen Durchschnittswertkostenfunktion $K = \frac{1}{N} \sum_{k=1}^N (x_k^2 + \vartheta^2 u_{k-1}^2)$ kann man zeigen, daß schon nach wenigen Perioden der asymptotische Wert $\lim_{N \rightarrow \infty} E\{K\} = E\{x^2 + \vartheta^2 u^2\} = \sigma_{xx}^2 + \vartheta^2 \cdot \sigma_{uu}^2$ erreicht wird, wobei σ_{xx}^2 und σ_{uu}^2 respektive die Varianzen von Zustand und Steuerung sind [5]). An f sind, wie wir sogleich sehen werden, noch weitere allerdings nur schwache Forderungen zu stellen.

Der Beweis für die Existenz dynamischer Sicherheitsäquivalente kann nun sehr rasch geführt werden.

Aufgrund von V2 und V1 folgt, daß Zustand und Steuerung ebenfalls normalverteilt sind, so daß

$$F := E \{ f(x,u) \} = F(\sigma_{xx}^2, \sigma_{uu}^2, \sigma_{xu}^2, \mu_x, \mu_u) \quad (3.2)$$

lediglich von den ersten und zweiten (zentrierten) Momenten der Zustands- und Steuervariablen abhängt. Ohne wesentliche Beschränkung der Allgemeinheit wollen wir der Einfachheit halber annehmen, daß F nicht von den Mittelwerten μ_x und μ_u und der Kovarianz σ_{xu}^2 abhängt. Häufig wird F von der Kovarianz sowieso nicht abhängen, da Zustand und Steuerung in der Kostenfunktion separiert auftreten, d.h.

$$f(x,u) = L(x) + P(u). \quad (3.3)$$

Damit wird aus (3.2)

$$F = F(\sigma_{xx}^2, \sigma_{uu}^2). \quad (3.4)$$

Eine notwendige Bedingung für die Existenz eines Minimums von F ist nun das Verschwinden der 1. Variation von F (nach der die lineare Politik bestimmenden Gewichtsfolge)

$$\delta F = \left. \frac{\partial F}{\partial \sigma_{xx}^2} \right|_0 \delta \sigma_{xx}^2 + \left. \frac{\partial F}{\partial \sigma_{uu}^2} \right|_0 \delta \sigma_{uu}^2 = 0 \quad (3.5)$$

Hierbei wird vorausgesetzt, daß die Ableitungen $\frac{\partial F}{\partial \sigma_{xx}^2}$ und $\frac{\partial F}{\partial \sigma_{uu}^2}$ existieren. Ferner wollen wir fordern, daß

$\frac{\partial F}{\partial \sigma_{xx}^2} > 0$ und $\frac{\partial F}{\partial \sigma_{uu}^2} > 0$ und beschränkt ist, so daß der Quotient

$$\theta^2 := \left. \frac{\frac{\partial F}{\partial \sigma_{uu}^2}}{\frac{\partial F}{\partial \sigma_{xx}^2}} \right|_0 \quad (3.6)$$

positiv und beschränkt ist.

Man beachte, daß die Forderungen nach Existenz der Ableitungen von F nach den Varianzen weit weniger einschränkend ist als Malinvauds Forderung nach zweimaliger Differenzierbarkeit von f nach dem Zustand. Unstetigkeiten in f sind dabei durchaus zugelassen. Dadurch wird es möglich, für f auch solche Funktionen zu verwenden, die geeignet sind, mögliche Beschränkungen im Zustands- und/oder Entscheidungsraum zu beschreiben. Die Positivität der Ableitungen dürfte in den allermeisten Fällen aus ökonomischen Gründen sichergestellt sein.

Mit θ^2 wird nun aus (3.5)

$$\delta \sigma_{xx}^2 + \theta^2 \delta \sigma_{uu}^2 = 0 \quad (3.7)$$

Dieser Ausdruck stellt ein bemerkenswertes Ergebnis dar. Er ist nämlich nichts anderes als die erste Variation eines quadratischen Kriteriums mit dem "reduzierten Funktional" ([6], Kap. 7)

$$F_Q := \sigma_{xx}^2 + \theta^2 \sigma_{uu}^2 \quad (3.8)$$

Damit ist die Existenz dynamischer Sicherheitsäquivalente aber fast schon bewiesen. Denn für ein durch (2.1), V2 und (3.8) definiertes Optimierungsproblem weist man leicht nach [6], daß die optimale Politik gegeben ist durch

$$u_k^* = \eta \left[\gamma x_k + \sum_{k'=0}^{\infty} z_1^{k'}(\theta) \hat{r}(k+1+k|k...) \right], \quad (3.9)$$

wo η, γ Konstanten sind,

$$z_1(\theta) := \frac{1 + 2\theta^2 - \sqrt{1 + 4\theta^2}}{2\theta^2} < 1$$

ist und $\hat{r}(k+1+k|k...)$ die durch (2.10) definierten dynamischen Sicherheitsäquivalente darstellen. D.h. auch bei nichtquadratischem Kriterium existieren unter den gegebenen Voraussetzungen dynamische Sicherheitsäquivalente, wobei man sich θ^2 lediglich über (3.6) zu verschaffen hat. Treten dabei mehrere θ^2 -Werte auf, so hat man dasjenige $\theta^2 > 0$ zu verwenden, das auf die niedrigsten Kosten führt.

Wir fassen das Ergebnis noch einmal zusammen in dem folgenden

Satz:

Unter den Voraussetzungen V1 bis V3 und der zusätzlichen Bedingung, daß die Ableitungen von F nach σ_{xx}^2 und σ_{uu}^2 positiv und (bzgl. σ_{uu}^2) beschränkt sind, existieren dynamische Sicherheitsäquivalente.

Hierbei ist die einzige Voraussetzung, die wirklich kritisch ist, die Forderung V1 nach Linearität der Politik.

Für quadratische Kriterien ist allerdings, wie man (2.4) entnimmt, diese Linearitätsvoraussetzung erfüllt. Es ist daher anzunehmen, daß für eine große Klasse nichtquadratischer Kriterien eine lineare Politik zumindest eine gute Approximation darstellt. Somit erwies sich z.B. das häufig unreflektiert angewandte Verfahren der Nachfrageprognostizierung als in vielen Fällen durchaus gangbar.

Literatur

- [1] Box und Jenkins; Time Series Analysis, Forecasting and Control, Holden-Day, 1970
- [2] Malinvaud, E.; First Order Certainty Equivalence; Econometrica, Vol. 37, 1969
- [3] Midler, J. L.; Optimal Control of a Discrete Time Stochastic System Linear in the State, J. Math. An. a. Appl., 25, 1969
- [4] Schneeweiß, Ch.; Optimale Prognosen und suffiziente Statistiken in quadratischen dynamischen Optimierungsproblemen, Stat. Hefte 1972
- [5] —————; Über den Zusammenhang von quadratischer stochastischer Dynamischer Programmierung und Wiener-Newton-Theorie; Operations Research Verfahren, XII, 1972
- [6] —————; Regelungstechnische stochastische Optimierungsverfahren, in der Reihe: Lecture Notes in Operations Research and Mathematical Systems, Springer-Verlag, Heidelberg, 1971
- [7] Theil, H.; A Note on Certainty Equivalence in Dynamic Planning, Econometrica, Vol. 25, 1957
- [8] —————; Optimal Decision Rules for Government and Industry; North-Holland Pub. Comp.; Amsterdam, 1964
- [9] Ziemba, W. T.; Transforming Stochastic Dynamic Programming Problems into Nonlinear Programs, Man. Sci., 17, 1971

Stochastische Programmierungsmodelle
als Vektormaximumprobleme
von W. Dürr, Regensburg

1. Das Vektormaximumproblem

In der Mathematischen Programmierung sind die Möglichkeiten der Optimierung genau einer Zielfunktion die hauptsächlichen Forschungsziele. In den Anwendungsgebieten der Mathematischen Programmierung, z.B. in der Betriebswirtschaftslehre, ist fast ausschließlich das Gewinnmaximierungsprinzip die Grundlage optimaler Entscheidungen. In neuerer Zeit hat das Problem der mehrfachen Zielsetzung (Zielkonflikt) oder Vektormaximumproblem zunehmend Beachtung gefunden.¹

Definition 1:

Es sei $X \subseteq \mathbb{R}^n$; $z_k(\underline{x})$ seien K Zielfunktionen, also $z_k : \mathbb{R}^n \rightarrow \mathbb{R}$ ($k = 1, \dots, K$). Die Aufgabe

$$(VMP) \text{ "max" } \left\{ \underline{z}(\underline{x}) \mid \underline{x} \in X \right\} \quad \text{mit } \underline{z}(\underline{x}) = \begin{pmatrix} z_1(\underline{x}) \\ \vdots \\ z_K(\underline{x}) \end{pmatrix}$$

heißt Vektormaximumproblem.²

Die Anführungszeichen bringen zum Ausdruck, daß nicht nur eine Zielfunktion zu maximieren ist. Es ergibt sich die Frage, welche Lösungen $\underline{x} \in X$ als "optimal" bezeichnet werden können. Dies sind sicherlich nur solche \underline{x} , die gemäß der folgenden Definition (funktional) effizient sind.

Definition 2:

$\underline{x} \in X$ heißt genau dann (funktional) effizient bezüglich X und $z_1(\underline{x}), \dots, z_K(\underline{x})$, wenn kein $\underline{x}' \in X$ mit der Eigenschaft

¹ Vgl. z.B. Kuhn-Tucker [1951], Karlin [1959], Bod [1963], Gutenberg [1966], Geoffrion [1967], Johnson [1968].

² Entsprechend wird ein Vektorminimumproblem definiert.

$$z_k(\underline{x}') \geq z_k(\underline{\hat{x}}) \quad (k = 1, \dots, K)$$

und

$$z_{k'}(\underline{x}') > z_{k'}(\underline{\hat{x}}) \quad \text{für mindestens ein } k' \in \{1, \dots, K\}$$

existiert.

Bei einem Entscheidungsproblem können nicht effiziente Lösungen vor vornherein außer Betracht bleiben. Der Entscheidungsträger wird sicherlich nur ein effizientes $\underline{x} \in X$ wählen, allerdings bleibt die Frage, welches effiziente \underline{x} er vorziehen wird. Er kann ein effizientes \underline{x} wählen, das Lösung eines "Kompromißprogramms" ist. Es bieten sich verschiedene Möglichkeiten für derartige Kompromißprogramme an¹, z.B.

$$(KP1) \quad \max \left\{ \sum_{k=1}^K t_k z_k(\underline{x}) \mid \underline{x} \in X \right\}$$

$$\text{mit } t_k \geq 0 \quad (k = 1, \dots, K).$$

$$(KP2) \quad \max \{ z_1(\underline{x}) \mid \underline{x} \in X \}$$

unter den Nebenbedingungen

$$\begin{aligned} z_2(\underline{x}) &\geq d_2 \\ &\vdots \\ z_K(\underline{x}) &\geq d_K. \end{aligned}$$

Die Frage, inwieweit Lösungen von (KP1) und (KP2) effizient sind, soll hier nicht diskutiert werden¹, desgleichen nicht, inwieweit die Menge aller effizienten Lösungen, die als "vollständige Lösung" von (VMP) bezeichnet sei, durch (KP1) und (KP2) bestimmt werden kann. Es sollen im folgenden nur gewisse Modelle der stochastischen Programmierung als Programme der Art (KP1) und (KP2) erkannt werden, was dann die Interpretation als Kompromißprogramme gewisser Vektormaximumprobleme erlaubt.

2. Stochastische Programmierung

Gegeben sei das lineare Programm

¹ Vgl. Geoffrion [1965], Dinkelbach [1971], Dürr [1971].

(LP) $\max \underline{c}x$
unter den Nebenbedingungen

$$\begin{aligned} \underline{A}x &\leq \underline{b} \\ \underline{0} &\leq x, \end{aligned}$$

wobei $\underline{c}; x \in \mathbb{R}^n; \underline{b} \in \mathbb{R}^m; \underline{A} \in \mathbb{R}^{m \times n}$.

Sind einige der Größen \underline{c} , \underline{b} , \underline{A} stochastisch, so verliert die obige Formulierung offensichtlich ihren Sinn. Ein Entscheidungsträger hat vorerst kein Kriterium für seine Entscheidungsfindung. Es werden üblicherweise Ersatzprogramme formuliert, wie zum Beispiel das Kompensationsprogramm, sowie das Chance-constrained Programm¹.

Beim Kompensations- oder auch Zweistufenprogramm werden Restriktionsverletzungen mit Strafkosten belegt, die dann von der Zielfunktion subtrahiert werden². Der Erwartungswert dieser Differenz wird dann maximiert. Es sei nun vorausgesetzt, daß nur \underline{b} stochastisch ist³. Dann lautet das Kompensationsprogramm wie folgt:⁴

$$\begin{aligned} \text{(EP1)} \quad \max (\underline{c}x + \sum_{i=1}^m q_i \int_{b_i < \underline{a}_i x} (b_i - \underline{a}_i x) f(b_i) db_i) &^5 \\ \text{unter der Nebenbedingung } x &\geq \underline{0}. \end{aligned}$$

Hierbei ist $q_i \geq 0$ ($i = 1, \dots, m$) der i -te Strafkostenkoeffizient, der die Bewertung der i -ten Restriktionsverletzung durch den Entscheidungsträger zum Ausdruck bringt, \underline{a}_i ($i = 1, \dots, m$) ist der i -te Zeilenvektor von \underline{A} . Die Zufallsvariable b_i möge die Dichtefunktion $f(b_i)$ besitzen ($i = 1, \dots, m$).

¹ Vgl. Kall [1966].

² Es sei angenommen, daß der Wert der Zielfunktion in Geldeinheiten gemessen wird.

³ Falls auch \underline{c} stochastisch ist, ergeben sich für das Folgende keine besonderen Schwierigkeiten.

⁴ Bezüglich einer allgemeineren Formulierung vgl. man Kall [1966], S. 255.

⁵ Die Integrale mögen selbstverständlich existieren.

Beim Chance-constrained Programm gilt es nach wie vor, die Zielfunktion \underline{cx} zu maximieren. Es wird nun weiter gefordert, daß die Nebenbedingungen mit gewissen vorgegebenen Mindestwahrscheinlichkeiten eingehalten werden müssen. Mit P_i ($i = 1, \dots, m$) als Wahrscheinlichkeitsmaß¹ lautet das Chance-constrained Programm also:

$$\begin{aligned} & \max \underline{cx} \\ & \text{unter den Nebenbedingungen} \\ (EP2) \quad & P_i(\{a_i x \leq b_i\}) = \int_{b_i > a_i x} f(b_i) db_i \geq \alpha_i; \quad 0 \leq \alpha_i \leq 1 \quad (i=1, \dots, m). \\ & \underline{x} \geq 0. \end{aligned}$$

Vergleicht man nun das Kompensationsprogramm (EP1) und das Chance-constrained Programm (EP2) mit (KP1) und (KP2), so stellt man fest, daß (EP1) und (EP2) nichts anderes sind als Kompromißprogramme für gewisse Vektormaximumprobleme. In der Tat hat man doch bei der Konzeption des Zweistufenmodells die mehrfache Zielsetzung, \underline{cx} sowie

$\int_{b_i > a_i x} (b_i - a_i x) f(b_i) db_i$ ($i = 1, \dots, m$) zu maximieren, bei der Konzeption

des Chance-constrained Modells hat man die mehrfache Zielsetzung, \underline{cx} sowie die Wahrscheinlichkeit der "Restriktionseinhaltung" $P_i(\{a_i x \leq b_i\})$ ($i = 1, \dots, m$) zu maximieren. Die obigen Überlegungen seien in dem folgenden Satz zusammengefaßt:

Satz 1:

Das Kompensationsprogramm (EP1) kann aufgefaßt werden als ein Kompromißprogramm (KP1) des Vektormaximumproblems

$$(VMP1) \text{ "max" } \left\{ \begin{array}{c} \underline{cx} \\ \int_{b_1 > a_1 x} (b_1 - a_1 x) f(b_1) db_1 \\ \vdots \\ \int_{b_m > a_m x} (b_m - a_m x) f(b_m) db_m \end{array} \right\} \quad \left. \begin{array}{c} \\ \\ \\ \end{array} \right\} \underline{x} \geq 0,$$

¹ $(\mathbb{R}_i, \mathcal{B}_i, P_i)$ ist der Wahrscheinlichkeitsraum von b_i , wobei \mathbb{R}_i die reellen Zahlen sind, und \mathcal{B}_i die zugehörige Borelalgebra.

das Chance-constrained Programm (EP2) kann aufgefaßt werden als ein Kompromißprogramm (KP2) des Vektormaximumproblems

$$(VMP2) \text{ "max" } \left\{ \begin{pmatrix} cx \\ P_1(\{a_1x \leq b_1\}) \\ \vdots \\ P_m(\{a_mx \leq b_m\}) \end{pmatrix} \middle| x \geq 0 \right\}.$$

Die Interpretation des Zweistufen- und des Chance-constrained Programms im Sinne von Vektormaximumproblemen führt zu der Frage, inwieweit effiziente Lösungen von (VMP1) auch effiziente Lösungen von (VMP2) sind und umgekehrt. Diese Frage wird im folgenden Abschnitt behandelt.

3. Vergleich der effizienten Lösungen von (VMP1) und (VMP2)

Ein Entscheidungsträger kann nicht erwarten, daß die Lösung des Kompensationsprogramms immer mit der Lösung des Chance-constrained Programms übereinstimmt, wenn er irgendwelche q_i beziehungsweise a_i ($i = 1, \dots, m$) wählt. Vielleicht ist es aber so, daß die "vollständigen Lösungen", also die Mengen der effizienten Lösungen von (VMP1) und (VMP2) übereinstimmen. Das ist in der Tat unter gewissen Voraussetzungen der Fall.

Satz 2:

Gilt fast überall $f(b_i) > 0$ ($i = 1, \dots, m$), so ist die "vollständige Lösung" von (VMP1) auch die "vollständige Lösung" von (VMP2), das heißt, jede effiziente Lösung von (VMP1) ist auch effiziente Lösung von (VMP2) und umgekehrt.

Beweis:

1. \hat{x} sei effiziente Lösung von (VMP2). Wäre \hat{x} nicht effiziente Lösung von (VMP1), gäbe es ein x_1 mit

$$a) \quad cx_1 > c\hat{x},$$

$$b_{i < a_i x_1} (b_i - a_i x_1) f(b_i) db_i \geq b_{i < a_i \hat{x}} (b_i - a_i \hat{x}) f(b_i) db_i \quad (i=1, \dots, m)$$

oder

b) $\underline{c}x_1 \geq \underline{c}\hat{x}$,

$$\int_{b_i < \underline{a}_i \hat{x}_1} (b_i - \underline{a}_i \hat{x}_1) f(b_i) db_i \geq \int_{b_i < \underline{a}_i \hat{x}} (b_i - \underline{a}_i \hat{x}) f(b_i) db_i \quad (i=1, \dots, m),$$

$$\int_{b_j < \underline{a}_j \hat{x}_1} (b_j - \underline{a}_j \hat{x}_1) f(b_j) db_j > \int_{b_j < \underline{a}_j \hat{x}} (b_j - \underline{a}_j \hat{x}) f(b_j) db_j$$

für mindestens ein $j \in \{1, \dots, m\}$.

Im Fall a) folgt sofort

$$\begin{aligned} \int_{b_i < \underline{a}_i \hat{x}} (b_i - \underline{a}_i \hat{x}_1) f(b_i) db_i + \int_{\underline{a}_i \hat{x}_1}^{\underline{a}_i \hat{x}_1} (b_i - \underline{a}_i \hat{x}_1) f(b_i) db_i &\geq \\ &\geq \int_{b_i < \underline{a}_i \hat{x}} (b_i - \underline{a}_i \hat{x}) f(b_i) db_i \quad (i=1, \dots, m). \end{aligned}$$

Da der zweite Summand auf der linken Seite stets kleiner oder gleich 0 ist, folgt

$$\int_{b_i < \underline{a}_i \hat{x}} (b_i - \underline{a}_i \hat{x}_1) f(b_i) db_i \geq \int_{b_i < \underline{a}_i \hat{x}} (b_i - \underline{a}_i \hat{x}) f(b_i) db_i \quad (i=1, \dots, m).$$

Daraus folgt aber $\underline{a}_i \hat{x}_1 \leq \underline{a}_i \hat{x}$ und schließlich

$$P_i(\{\underline{a}_i \hat{x}_1 \leq b_i\}) \geq P_i(\{\underline{a}_i \hat{x} \leq b_i\}) \quad (i = 1, \dots, m).$$

Mit $\underline{c}x_1 > \underline{c}\hat{x}$ folgt dann aber ein Widerspruch zur Effizienz von \hat{x} bezüglich (VMP2).

Im Fall b) folgt ebenso

$$\underline{a}_i \hat{x}_1 \leq \underline{a}_i \hat{x} \quad (i = 1, \dots, m) \text{ und}$$

$$\underline{a}_j \hat{x}_1 < \underline{a}_j \hat{x} \quad \text{für mindestens ein } j \in \{1, \dots, m\}.$$

Aufgrund der Voraussetzung folgt dann

$$P_i(\{\underline{a}_i \hat{x}_1 \leq b_i\}) \geq P_i(\{\underline{a}_i \hat{x} \leq b_i\}) \quad (i = 1, \dots, m) \text{ und}$$

$$P_j(\{\underline{a}_j \hat{x}_1 < b_j\}) > P_j(\{\underline{a}_j \hat{x} \leq b_j\}) \quad \text{für mindestens ein } j \in \{1, \dots, m\}.$$

Das ist aber ein Widerspruch zur Effizienz von \hat{x} bezüglich (VMP2).

2. Es sei nun \hat{x} effiziente Lösung von (VMP1). Wäre \hat{x} nicht effiziente Lösung von (VMP2), gäbe es ein x_1 mit

a) $\underline{c}x_1 > \underline{c}\hat{x}$,

$$P_i(\{a_i x_1 \leq b_i\}) \geq P_i(\{a_i \hat{x} \leq b_i\}) \quad (i = 1, \dots, m)$$

oder

b) $\underline{c}x_1 \geq \underline{c}\hat{x}$

$$P_i(\{a_i x_1 \leq b_i\}) \geq P_i(\{a_i \hat{x} \leq b_i\}) \quad (i = 1, \dots, m)$$

$$P_j(\{a_j x_1 \leq b_j\}) > P_j(\{a_j \hat{x} \leq b_j\}) \text{ für mindestens ein } j \in \{1, \dots, m\}.$$

Im Fall a) schließt man $a_i x_1 \leq a_i \hat{x}$ und daraus ähnlich wie oben

$$\int_{b_i < a_i x_1} (b_i - a_i x_1) f(b_i) db_i \geq \int_{b_i < a_i \hat{x}} (b_i - a_i \hat{x}) f(b_i) db_i$$

$$(i = 1, \dots, m).$$

Mit $\underline{c}x_1 > \underline{c}\hat{x}$ ist das aber ein Widerspruch zur Effizienz von \hat{x} bezüglich (VMP1).

Im Fall b) schließt man wiederum

$a_i x_1 \leq a_i \hat{x}$ ($i = 1, \dots, m$) und aufgrund der Voraussetzung des Satzes $a_j x_1 < a_j \hat{x}$ für mindestens ein $j \in \{1, \dots, m\}$.

Das hat aber zur Folge

$$\int_{b_i < a_i x_1} (b_i - a_i x_1) f(b_i) db_i \geq \int_{b_i < a_i \hat{x}} (b_i - a_i \hat{x}) f(b_i) db_i \quad (i=1, \dots, m),$$

$$\int_{b_j < a_j x_1} (b_j - a_j x_1) f(b_j) db_j > \int_{b_j < a_j \hat{x}} (b_j - a_j \hat{x}) f(b_j) db_j$$

$$\text{für mindestens ein } j \in \{1, \dots, m\}.$$

Zusammen mit $\underline{c}x_1 \geq \underline{c}\hat{x}$ bedeutet dies einen Widerspruch zur Effizienz von \hat{x} bezüglich (VMP1).

Damit ist Satz 2 bewiesen.

L <

Satz 2 läßt die folgende Auslegung zu: Interpretiert man das Zweistufenmodell und das Chance-constrained Modell als Vektormaximumprobleme, so sind sie unter den Voraussetzungen des Satzes äquivalent, da sie zu denselben effizienten Lösungen führen. Zu jedem q_i gibt es also ein α_i ($i=1, \dots, m$), so daß das Zweistufen- und Chance-constrained Programm die gleiche effiziente Lösung liefern. Es besteht keine Veranlassung, die Konzeption des einen Modells der Konzeption des anderen Modells vorzuziehen.

Sind die Voraussetzungen von Satz 2 nicht erfüllt, so lassen sich leicht Punkte angeben, die bezüglich des einen Vektormaximumproblems effizient sind, bezüglich des anderen aber nicht, das heißt aber, ein optimaler Punkt¹ des Zweistufenmodells braucht nicht optimaler Punkt des Chance-constrained Modells zu sein und umgekehrt.²

Beispiel:

$$\begin{array}{ll}
 \max (x_1 + x_2) \\
 \text{unter den Nebenbedingungen} \\
 \frac{1}{2}x_1 + \frac{1}{2}x_2 \leq b_1 \\
 \frac{2}{3}x_1 + \frac{1}{3}x_2 \leq b_2 \\
 x_1, x_2 \geq 0
 \end{array}
 \quad
 f(b_2)=f(b_1) = \begin{cases} 0 & \text{für } -\infty < b_1, b_2 < 0 \\
 1 & \text{" } 0 \leq b_1, b_2 \leq \frac{1}{2} \\
 0 & \text{" } \frac{1}{2} \leq b_1, b_2 < 1 \\
 \frac{1}{2} & \text{" } 1 \leq b_1, b_2 \leq 2 \\
 0 & \text{" } 2 < b_1, b_2 < \infty
 \end{cases}$$

Das Chance-constrained Programm mit $\alpha_1 = \alpha_2 = \frac{1}{2}$ lautet dann:

$$\begin{array}{ll}
 \max (x_1 + x_2) \\
 \text{unter den Nebenbedingungen} \\
 P_1(\{\frac{1}{2}x_1 + \frac{1}{2}x_2 \leq b_1\}) \geq \frac{1}{2} \\
 P_2(\{\frac{2}{3}x_1 + \frac{1}{3}x_2 \leq b_2\}) \geq \frac{1}{2}
 \end{array}$$

¹ Im Sinne des Effizienzbegriffs.

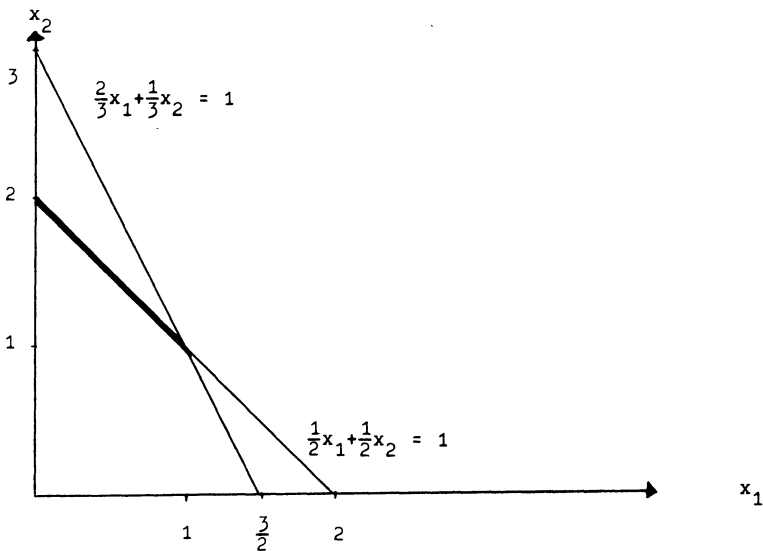
² Unter gewissen Voraussetzungen läßt sich aber zeigen, daß es zumindest gemeinsame effiziente Punkte gibt. Vgl. das folgende Beispiel.

Lösungen und effizient¹ bezüglich des Vektormaximumproblems (VMP2) dieses Beispiels sind alle Punkte, die auf der Verbindungsstrecke der Punkte (1,1) und (0,2) liegen. Für diese Punkte gilt

$$\frac{1}{2}x_1 + \frac{1}{2}x_2 = 1$$

$$\frac{2}{3} \leq \frac{2}{3}x_1 + \frac{1}{3}x_2 \leq 1$$

Der Wert der Zielfunktion ist $x_1 + x_2 = 2$.



Lediglich der Punkte (0,2) ist aber auch effizient bezüglich des Vektormaximumproblems (VMP1) dieses Beispiels, denn für alle Punkte, die auf der Verbindungsstrecke der Punkte (1,1) und (0,2) liegen, sind die

¹ Das heißt nicht, daß immer alle Lösungen von (KP2) effizient sein müssen. Vgl. S. 2.

Werte der ersten beiden Zielfunktionen $x_1 + x_2$ und

$\int_{b_1 < \frac{1}{2}x_1 + \frac{1}{2}x_2} (b_1 - (\frac{1}{2}x_1 + \frac{1}{2}x_2))f(b_1)db_1$ konstant, während die dritte Zielfunktion

$\int_{b_2 < \frac{2}{3}x_1 + \frac{1}{3}x_2} (b_2 - (\frac{2}{3}x_1 + \frac{1}{3}x_2))f(b_2)db_2$ für $(0,2)$ maximal wird. Für den Punkt

$(0,2)$ ergibt sich

$$\int_{b_2 < \frac{2}{3}} (b_2 - \frac{2}{3})f(b_2)db_2 = \frac{1}{2} \int_0^{\frac{2}{3}} (b_2 - \frac{2}{3})db_2 = -\frac{5}{24}.$$

Für den Punkt $(1,1)$ ergibt sich

$$\int_{b_2 < 1} (b_2 - 1)f(b_2)db_2 = \frac{1}{2} \int_0^1 (b_2 - 1)db_2 = -\frac{9}{24}.$$

Literaturverzeichnis

BOD, P.: Lineare Optimierung mittels simultan gegebener Zielfunktionen.

In: A. PREKOPA (ed.), Colloquium on Applications of Mathematics to Economics, Budapest 1963, 55-60. Budapest: Akadémiai Kiadó 1965.

DINKELBACH, W.: Über einen Lösungsansatz zum Vektormaximumproblem.

"Neuere Ergebnisse der Unternehmensforschung" in den Lecture Notes (50) des Springer-Verlages 1971.

DÜRR, W.: Einige Theoreme zum Vektormaximumproblem. In: Regensburger Diskussionsbeiträge zur Wirtschaftswissenschaft, Serie B, Nr. 1. Regensburg 1971.

GEOFFRION, A.M.: A parametric programming solution to the vector maximum problem, with applications to decisions under uncertainty. Stanford, Cal.: Graduate School of Business, Tech. Report No. 11, 1965 (Clearinghouse: AD 613-670).

GEOFFRION, A.M.: Solving bicriterion mathematical programs. OR 15, 39-54 (1967).

- GUTENBERG, E.: Über einige Fragen der neueren Betriebswirtschaftslehre. Zeitschrift für Betriebswirtschaftslehre, Ergänzungsheft 1, 1-17 (1966).
- JOHNSON, E.: Studies in multiobjective decision models. Lund: Studentent-literatur 1968.
- KALL, P.: Qualitative Aussagen zu einigen Problemen der stochastischen Programmierung. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 6, 246-272 (1966).
- KARLIN, S.: Mathematical methods and theory in games, programming, and economics, Vol. I, Reading, Mass.: Addison-Wesley 1959.
- KUHN, H.W., and TUCKER, A.W.: Nonlinear programming. In: J. NEYMAN (ed.), Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 481-492. Berkeley, Cal.: University of California Press 1951.

Langfristige Planung

**Ein Modell für die strategische Zielsetzung
und Ressourcen-Zuteilung in divisional gegliederten Unternehmen**
von L. Peters, Düsseldorf

I. EINLEITUNG

Bei der Festlegung der Unternehmensziele steht die Unternehmensleitung vor der Aufgabe, einen Kompromiss zwischen den Anforderungen der Unternehmensleitung und den verschiedenen Interessengruppen zu finden. Die Unternehmensleitung plant, um die Stellung des Unternehmens im Markt halten und weiter ausbauen zu können, ein bestimmtes Umsatz- und Investitionsvolumen bei möglichst hoher Rentabilität des eingesetzten Kapitals. Die Kapitaleigner fordern eine angemessene Gewinnausschüttung, und die Kreditinstitute achten bei der Vergabe ihrer Mittel auf die Einhaltung bestimmter Mindestkonditionen in der Kreditwürdigkeit.

Wenn ein Konzern nach ergebnisverantwortlichen Geschäftsbereichen gegliedert ist, muss die Unternehmensleitung ausserdem noch für einen Kompromiss zwischen den Zielvorstellungen des Gesamtkonzerns und der einzelnen Geschäftsbereiche sorgen.

Es müssen also primäre Gesamtziele des Konzerns, wie z.B. Rendite und das zugehörige Risiko, in Einklang gebracht werden mit den gleichen primären Einzelzielen der Geschäftsbereiche unter Beachtung sekundärer Gesamtziele des Konzerns, wie z.B. Umsatzwachstum, Investitionsvolumen, Verschuldungsgrad und Gewinnausschüttung. Bei der Abstimmung zwischen Konzern- und Geschäftsbereichszielen sind zwei Punkte zu beachten:

- ¶ Die von den Geschäftsbereichen angeforderten Investitionsmittel werden gewöhnlich die verfügbaren Mittel überschreiten. Die knappen Ressourcen müssen daher so verteilt werden, dass die Gesamtziele mit der Summe der Einzelziele übereinstimmen.

- ¶ Aus den Abhängigkeiten zwischen Geschäftsbereichen ergeben sich Rückwirkungen auf Rendite- und Risikoziele des Konzerns, aus denen sich im Vergleich zu unabhängig voneinander betrachteten Geschäftsbereichen auch unterschiedliche Einzelziele für die Geschäftsbereiche ableiten lassen.

Die Wahl bestimmter Ziele wird bei den auftretenden Unsicherheiten in den Projektionen der Geschäftstätigkeiten von drei Risikokomponenten beeinflusst: (a) Abweichung (Varianz) vom Erwartungswert des Erfolgsmassstabes (Gewinn, Rendite) der Zielsetzung, (b) Unterschiedliche Abhängigkeitsbeziehungen (Kovarianzen) zwischen einzelnen Bestandteilen eines Erfolgsmassstabes und (c) Risikopräferenzen der Interessengruppen.

In diesem Beitrag soll ein Modell diskutiert werden, das dem Management aussagefähige Entscheidungsgrundlagen für die Festlegung der Ziele liefert. Dabei handelt es sich um ein erweitertes Portfolio-Selection-Modell, das die wesentlichen Variablen bei der Zielsetzung transparent macht und so ein wirksames Kommunikationsmittel für alle am Zielsetzungsprozess Beteiligten ist. Die Betrachtung der Geschäftsbereiche als ein Portefeuille hat für die Unternehmensleitung zwei wichtige Konsequenzen:

- ¶ Vermeintlich hohes Risiko bei Betrachtung der Geschäftsbereiche als voneinander unabhängig kann sich als akzeptables niedriges Risiko im Portefeuille herausstellen (oder umgekehrt).
- ¶ Alternativ effiziente Portefeuille-Zusammensetzungen stellen das Management vor die Wahl einer bestimmten Zielgrössenkombination, ohne die explizite Angabe der jeweiligen Risikopräferenz zu verlangen.

Im folgenden wird der Ablauf des Zielsetzungsprozesses dargestellt. Im Anschluss daran werden die Struktur des Modells vorgestellt und seine praktische Anwendung an einem Beispiel erläutert.

II. ZIELSETZUNGSPROZESS

Gesamtziele für den Konzern und Einzelziele für die Geschäftsbereiche werden in einem Prozess abgeleitet, an dem hauptsächlich drei Parteien beteiligt sind:

- (a) die zentrale Unternehmensplanung, (b) die einzelnen Geschäftsbereiche und
- (c) die Unternehmensleitung.

Die zentrale Unternehmensplanung untersucht zunächst unabhängig von anderen Stellen des Konzerns Entwicklungen der Gesamtkonjunktur, des Wettbewerbs und der Position des Konzerns im Markt, um für die wichtigsten globalen Ziele des Konzerns einen Orientierungsrahmen aufzustellen. Dabei werden Erwartungswerte und eine grobe Schätzung der Bandbreite ihrer Eintrittswahrscheinlichkeit angegeben. Die Geschäftsbereiche erarbeiten unabhängig von den Globalzielen Vorschläge für ihre eigene Entwicklung. Diese Teilanalysen ermitteln Prognosewerte für die wichtigsten Erfolgsfaktoren, die unter Berücksichtigung der eigenen Aktionsmöglichkeiten zu ersten Zielvorstellungen korrigiert werden. Die Unternehmensplanung konsolidiert die Vorschläge der Geschäftsbereiche und stellt die ermittelten Gesamtzahlen den ursprünglichen Konzernzielen gegenüber. Grobe Diskrepanzen werden mit den Geschäftsbereichen abgestimmt. Die Konzernleitung beurteilt die modifizierten Vorschläge für Gesamt- und Einzelziele aus übergeordneter Konzernsicht und setzt möglicherweise neue Schwerpunkte durch eine veränderte Zuteilung der knappen Ressourcen. Die abschliessende Diskussion zwischen der Konzernleitung, den Geschäftsbereichen und der Unternehmensplanung setzt die endgültigen Ziele fest, aufgrund derer die Geschäftsbereiche ihre strategischen Pläne ausarbeiten können.

Für analytische Zwecke benutzt die Unternehmensplanung in den einzelnen Schritten des Abstimmungsprozesses eine Reihe von Modellen (Schaubild 1):

- 1 Zentral ermittelte Konzernziele werden in Alternativen durchgespielt, um ihre Auswirkungen auf Bilanz, GuV und Cash-Flow-Rechnung transparent zu machen. Für die wichtigsten Grössen wird gleichzeitig das Risiko analysiert.

- ¶ Vorschläge der Geschäftsbereiche werden ebenfalls in Alternativen durchgerechnet, um Änderungen bestimmter Annahmen in ihren Auswirkungen auf einzelne Geschäftsbereiche und auf die jeweiligen Summen der einzelnen Ziele testen zu können. Die wesentlichsten Grössen werden auch hier einer Risikoanalyse unterzogen.
- ¶ Das im nächsten Kapitel vorgestellte Modell betrachtet die Vorschläge der Geschäftsbereiche schliesslich als ein Portefeuille, in dem die Abhängigkeitsbeziehungen zwischen Geschäftsbereichen durch Korrelierungen über gemeinsame übergeordnete Faktoren wie Preis- und Kostenindizes ausgedrückt werden, von denen die teilweise überlappenden Geschäftstätigkeiten in unterschiedlicher Weise abhängig sind. Der unterschiedlichen Risikobereitschaft von Interessengruppen trägt das Modell dadurch Rechnung, dass es der Konzernleitung nicht eine einzige Kombination von Gesamtzielen, sondern mehrere "effiziente" alternative Kombinationen zur endgültigen Auswahl vorschlägt.* Die quadratische Funktion des Modells berücksichtigt dabei die Beobachtung, dass die Risikobereitschaft der Interessengruppen in der Regel abnimmt, je mehr bei zusätzlichen Gewinnen zusätzliches Risiko in Kauf genommen wird.

Die aus der Simulation ermittelten Gesamtziele und das Portefeuille der vorgeschlagenen Einzelziele der Geschäftsbereiche können nur zufällig übereinstimmen und müssen im Normalfall in den oben beschriebenen Abstimmungsschritten einander angepasst werden. Für eine bestimmte Zielgrösse, z.B. Rendite, muss das Gesamtziel nicht nur im Erwartungswert, sondern auch im Risiko mit der Summe der Einzelziele deckungsgleich sein (Schaubild 2). Bei der Anpassung lassen sich drei Hauptfälle unterscheiden:

* - Eine Gewinn-Risiko-Kombination ist effizient, wenn für einen bestimmten Gewinn minimales Risiko oder für ein bestimmtes Risiko maximaler Gewinn erreicht ist.

1. Der Erwartungswert für die Summe der Geschäftsbereichsziele liegt für ein spezifisches Ziel, wie z. B. die Rendite, bei gleichem Risiko höher (niedriger) als das Gesamtziel. Entweder war das Gesamtziel zu pessimistisch (optimistisch) und wird angepasst, oder die Konzernleitung unterstellt den Geschäftsbereichen zu viel Optimismus (Pessimismus) und fordert realistischere Vorschläge.

2. Das Risiko für die Summe der Geschäftsbereichsziele ist bei gleicher Renditeerwartung grösser (kleiner) als das Risiko des Gesamtziels. Informationen und Annahmen bei Ableitung der Risikowerte werden nochmals überprüft, oder die Unterschiede erklären sich aus der Methodik der Datenaufbereitung (z. B. mit und ohne Berücksichtigung von Abhängigkeiten der Geschäftsbereiche untereinander).

3. Gesamtziele und Summe der Einzelziele stimmen weder in der Rendite noch im Risiko überein und eine Kombination der beschriebenen Massnahmen ist erforderlich.

Während sich die Kurve effizienter Kombinationen von Rendite und Risiko für die Geschäftsbereichsziele durch Veränderung der Risikopräferenz bestimmen lässt, ist eine eindeutige Aussage über die Effizienz der Konzernziele im Vergleich zu den Geschäftsbereichszielen nicht möglich. Denn Konzernzielen und Geschäftsbereichszielen liegen in der Regel unterschiedliche Informationen und Annahmen zugrunde, so dass mehrere Kurven effizienter Kombinationen denkbar sind. In der Praxis spielt die Frage der Zieleffizienz ohnehin nur eine untergeordnete Rolle. Der provozierte Konflikt zwischen Gesamt- und Einzelzielen regt vielmehr an, den Ursachen der Diskrepanz auf die Spur zu kommen und dann je nach dem Anspruchsniveau die Abstimmung herbeizuführen.

III. STRUKTUR DES MODELLS

Grundlage des Modells ist das bekannte Portfolio-Selection-Problem [1], das in drei Richtungen erweitert ist:

1. Maximiert wird der Erwartungsnutzen (Bernoulli-Prinzip) bzw. das Sicherheitsäquivalent.
2. Die Variablen korrelieren nicht direkt miteinander, sondern indirekt über gemeinsame Indizes (Sharpe-Modell) [3].
3. Simultan gekoppelt mit den Portefeuille-Variablen x ist das Investitionsvolumen y und die Fremdmittelaufnahme z .

Das mehrperiodige Modell hat folgende Grundstruktur:

$$\max \left\{ g'x - \lambda x' B x - a'y - c'z \mid x \leq d, y = \chi, x \geq 0, y, z \geq 0 \right\}$$

mit weiteren Nebenbedingungen für die Korrelierung der Variablen x über gemeinsame Indizes und für die Einhaltung des finanziellen Gleichgewichts. In der Zielfunktion ist x das zusätzliche Umsatzvolumen pro Geschäftsbereich, g der zusätzliche Bruttogewinnbeitrag pro 1 DM zusätzlichen Umsatz, y das zusätzliche Investitionsvolumen, a die für 1 DM zusätzlichen Investitionsvolumens erforderlichen Fixkosten, z die zusätzlich benötigten Kredite, c die pro 1 DM zusätzlicher Kredite benötigten Kapitalkosten, d das erwartete, absolute Umsatzwachstum der Geschäftsbereiche, χ die marginalen Kapitalumschlagshäufigkeiten, λ der Risikoeffizient und B die Matrix der Varianzen mit Nicht-Null-Elementen nur in der Diagonalen.

Ein wesentliches Merkmal des Modells ist die Kopplung des Investitionsvolumens an das Umsatzwachstum über marginale Kapitalumschlagshäufigkeiten. Im Gegensatz zu kurzfristigen Entscheidungsproblemen, in denen konkrete Entscheidungsalternativen in der Regel bekannt sind, besteht eines der Hauptprobleme der langfristigen Planung in der Vorhersage konkreter Aktivitätsmöglichkeiten. Investitionsprojekte lassen sich nur in Ausnahmefällen (z. B. Raffinerien) fünf und mehr Jahre voraus-

planen. Die noch am besten zu planende Grösse ist der Absatz, den die Geschäftsbereiche aufgrund von Verbrauchstrends, bekannten und vermuteten Kapazitätserweiterungen der Konkurrenz sowie eigenen Aktionsprogrammen angeben können. Ist die Kapitalumschlagshäufigkeit bekannt, liesse sich aus den technischen Kapazitätserfordernissen des zusätzlichen Absatzes das zusätzliche Investitionsvolumen ableiten. Um Kapitalumschlagshäufigkeiten für ganze Geschäftsbereiche zu ermitteln, müssen die mengenmässigen Absatzausdehnungen einzelner Produktgruppen zu einem wertmässigen Umsatzvolumen zusammengefasst werden. Aus verschiedenen Gründen wird sich über eine längere Periode das Verhältnis von zusätzlichem Umsatz zu den Investitionen verändern: Konjunkturschwankungen haben unterschiedliche Auswirkungen auf Preise und Kosten der Absatzprodukte und Investitionsgüter, die Marktnähe einzelner Produkte verschiebt sich, zwischen Umsatzzuwachs und Realisierung der Investitionen liegen Zeitverzögerungen, die Kapazitätsrelevanz (Umweltschutz!) von Sachanlagen ändert sich usw. Da der zusätzliche Umsatz neues Kapital nicht nur für Sachanlagen, sondern in einem bestimmten Verhältnis auch für Umlaufvermögen erfordert, werden die Kapitalumschlagshäufigkeiten für Sachanlagen und Umlaufvermögen getrennt.

Die Gewinnentwicklung der Geschäftsbereiche wird von der Entwicklung ausgewählter Indizes, z. B. Preis- und Kostenindizes der Industrie, abhängig gemacht. Jeder Geschäftsbereich wird aufgrund spezifischer Entwicklungen seiner Teilmärkte und verschiedener Kostenentwicklungen durch Änderungen des Beschäftigungsgrads und der Verfahren in unterschiedlichem Masse von diesen Globalindizes beeinflusst. Während die Stäbe der Geschäftsbereiche die Parameter für die Abhängigkeit von der Index-Entwicklung und die möglichen Abweichungen vom erwarteten Abhängigkeitsverhältnis schätzen, ist es Aufgabe der Unternehmensplanung, Erwartungswerte und Varianzen von Änderungsraten der Indizes zu prognostizieren. Da der gesamte Planungshorizont zweckmässigerweise in einzelne Teilperioden aufgeteilt wird, können die einzelnen Ereignisse der Wahrscheinlichkeitsverteilungen in den verschiedenen Perioden über konditionale Wahrscheinlichkeiten miteinander verbunden werden.

Begrenzt werden die Möglichkeiten der Expansion durch den finanziellen Spielraum. Die Mittelverwendung in Sachanlagen und Umlaufvermögen darf in keiner Periode die aus einbehaltenen Gewinnen, Abschreibungen, Kapitalerhöhungen und neu aufgenommenen Krediten zur Verfügung stehenden Finanzmittel überschreiten (unter Berücksichtigung bestimmter Dividendensätze, unterschiedlicher Steuersätze usw.). Je nach Detaillierungsgrad des Modells lassen sich Sekundärziele des Konzerns in die Nebenbedingungen als Mindestanforderungen aufnehmen, z. B. Einhaltung einer Verschuldungsgrenze oder Erreichen einer Mindest-Gesamtkapitalrendite. Werden diese Ziele nicht direkt ins Modell aufgenommen, müssen mehrere Versuchsläufe eine Abstimmung herbeiführen. Die schrittweise Änderung der Risikoneigung erzeugt alternativ effiziente Gewinn / Risiko-Kombinationen für die Geschäftsbereiche, aus denen die Konzernleitung unter Beachtung des jeweiligen Niveaus der Sekundärziele eine bestimmte Kombination auswählt.

IV. BEISPIEL "ABC-CHEMIE"

Das Chemie- Unternehmen "ABC-Chemie" operiert in drei Geschäftsbereichen: Organika, Kunststoffe, Pharmazeutika. Organika ist über die Produktgruppe Weichmacher eng mit den Kunststoffen verbunden, so dass Preise und Kosten beider Geschäftsbereiche relativ stark korrelieren. Pharmazeutika arbeitet völlig unabhängig auf einem eigenen Markt.

ERGEBNIS DER VORANALYSEN

Die Einzelanalysen der Geschäftsbereiche und der Unternehmensplanung für die Zielsetzung 1972-76 haben folgende Ergebnisse gebracht (alles fiktive Zahlen):

- § In Preisen von 1971 lässt sich bei einer durchschnittlichen Wachstumsrate der chemischen Industrie von 11% in den Geschäftsbereichen Pharmazeutika und Organika ein überdurchschnittliches Umsatzwachstum erwarten. Kunststoffe liegen unter dem Durchschnitt (Schaubild 3).

¶ Die Prognose marginaler Kapitalumschlagshäufigkeiten (Sachanlagen) zeigt zunehmende Kapitalintensität bei Organika und Kunststoffen (Schaubild 4).

¶ Für die gesamte chemische Industrie wird mit einem weiteren Preisverfall in den Jahren 1972-1973 von ca. 4 Punkten und einer leichten Erhöhung von 2 Punkten in den Jahren bis 1976 gerechnet. Gleichzeitig werden in den ersten zwei Jahren weitere Kostensteigerungen in Höhe von ca. 16% und in den folgenden drei Jahren in Höhe von durchschnittlich ca. 21% erwartet (Schaubild 5).

¶ Aufgrund unterschiedlicher Abhängigkeiten der Geschäftsbereiche von den Preis- und Kostenindizes der chemischen Industrie und aufgrund unterschiedlicher Produktivitätssteigerungen durch neue Produktionsverfahren können die Kostensteigerungen bei den nur wenig schwankenden Preisen teilweise zwar aufgefangen werden, in den vom Umsatzvolumen her wichtigsten Geschäftsbereichen Organika und Kunststoffe jedoch drohen sinkende Gewinne. Allein Pharmazeutika zeigen eher eine negative Korrelation mit der Entwicklung der Gesamtindustrie, die sich in steigenden - aber auch mit grösserem Risiko behafteten - Gewinnerwartungen ausdrückt (Schaubild 6).

Unabhängig von den Projektionen der Geschäftsbereiche hat die zentrale Unternehmensplanung in gründlichen Analysen des Wettbewerbs und der eigenen Position im Wettbewerb folgende Konzernziele ermittelt:

	<u>1972-73</u>	<u>1974-76</u>
1. Gesamtkapitalrendite	16%	13%
2. Umsatzwachstum Ø	12%	14%
3. Investitionsvolumen	450 Mio.	1100 Mio.
4. Verschuldung (Eigenkapital/Gesamtkapital)	.65	.60

In der Risikoanalyse wurde die Wahrscheinlichkeit einer Abweichung von ± 2 Prozentpunkten der Gesamtkapitalrendite mit 50% ermittelt.

Die Konzernleitung hält den Rückgang der ohnehin nicht überragenden Rendite (vor Steuern!) bei einer Verschlechterung des Verschuldungsgrades und einem relativ hohen Risiko nicht für angemessen. Auf der anderen Seite befürchtet sie, bei zu geringem Wachstum die Marktposition zu verlieren. Sie fordert die Unternehmensplanung auf, Alternativen durchzuspielen, in denen durch eine günstigere Verteilung der Investitionssumme eine oder mehrere Zielgrößen verbessert werden.

Der Einfachheit halber wurden nur die zusätzlichen Umsätze, Investitionen und Fremdmittel als Variable ins Modell aufgenommen. Die Werte der optimalen Lösungen wurden dann zum jeweiligen Ausgangsniveau 1971 - das hier nicht im einzelnen aufgeführt ist - hinzugezählt, um vollständige Ergebnisrechnungen, Bilanzen und Cash-Flow-Rechnungen aufzustellen. Dabei galten die Annahmen, dass Ersatzinvestitionen in Höhe der Abschreibungen durchgeführt und Gewinne aus dem "alten" Umsatz für Dividenden auf das "alte" Grundkapital und für Sonderabschreibungen verwendet wurden. Das vom Modell ermittelte Investitionsvolumen besteht demnach nur aus Erweiterungsinvestitionen.

ERGEBNIS DES ZIELSETZUNGSPROZESSES

Die Feinabstimmung zwischen zentral ermittelten (ursprünglichen) Konzernzielen, Summe der einzelnen Geschäftsbereichsvorschläge und den Portefeuille-Vorschlägen des quadratischen Modells brachte folgende Resultate:

- ¶ Das ursprüngliche Konzernziel für den Bruttogewinn - Grundlage der Gesamtkapitalrendite von durchschnittlich 16% in den Jahren 1972-73 und von 13% für 1974-76 - hat sich im Vergleich zu den Modellvorschlägen als "ineffizient" erwiesen (Schaubild 7). Das Ziel wird als ineffizient bezeichnet, weil bei gleicher Gewinnerwartung eine Kombination von Geschäftsbereichsvorschlägen mit einem geringeren Risiko existiert. Das ursprüngliche Konzernziel basierte entweder auf anderen Annahmen für die Gesamtentwicklung als die Geschäftsbereichsvorschläge, oder den Abhängigkeiten zwischen den Geschäftsbereichen wurde nicht Rechnung getragen.

- ¶ Auch die Summe der einzelnen Vorschläge aus den Geschäftsbereichen weist bei gleicher Gewinnerwartung ein höheres Risiko aus als der Modellvorschlag (Schaubild 8). * In der Portefeuille-Betrachtung ergibt sich im Vergleich zur isolierten Addition der Einzelrisiken in der Regel ein anderes Risiko, weil bei Berücksichtigung der Abhängigkeiten zwischen den Geschäftsbereichen zusätzlich Kovarianzen zu beachten sind. Die Konzernleitung hätte die Summe der Einzelvorschläge als Konzernziele möglicherweise wegen scheinbar zu hohen Risikos abgelehnt. Das tatsächliche Risiko der Geschäftsbereiche bleibt dasselbe, ob einzeln oder als Portefeuille betrachtet: nur ist im Fall der Einzelbetrachtung die Informationsbasis unvollständig.
- ¶ Die Wahl bestimmter Einzelziele - und damit indirekt der Konzernziele - hängt u. a. von der Risikopräferenz der Konzernleitung ab bzw. von der Risikopräferenz, die die Konzernleitung anderen Interessengruppen unterstellt (Schaubild 9). Bei äusserster Risikoneigung ($\lambda = 1/55$) würden der Geschäftsbereich Organika mit einem Investitionsvolumen von 700 Mio. (Sachanlagen) einen Umsatz von 490 Mio. erzielen, Pharmazeutika mit 185 Mio. Investitionsvolumen einen Umsatz von 222 Mio. haben, Kunststoffe keine zusätzlichen Mittel bekommen und so keinen zusätzlichen Umsatz machen. Bei geringerer Risikoneigung (z. B. $\lambda = 1/5$) würden die Investitionen zunächst nur für Organika gekürzt, bei sehr starker Risikoaversion schliesslich auch für Pharmazeutika.
- ¶ Umgerechnet auf den Konzern ergeben die einzelnen Werte der optimalen Lösungen unterschiedliche Kombinationen der vier wesentlichsten Zielgrössen bei variierender Risikopräferenz. Dabei stellt

* - Alle folgenden Ergebnisse beziehen sich nur auf die Teilperiode 1974-76, um das darzustellende Zahlenmaterial zu reduzieren.

sich heraus, dass sich die ursprünglichen Konzernziele für 1976 nur bei äusserster Risikopräferenz und nur mit einer höheren Verschuldung als geplant realisieren lassen (Schaubild 10).

- § Ursprüngliche Konzernziele und die Summe der Geschäftsbereichsvorschläge lagen zwar beträchtlich auseinander. Die von der Konzernleitung als verbindlich verabschiedete Portefeuille-Lösung bei äusserster Risikoneigung ändert jedoch die ursprünglichen Konzernziele nur hinsichtlich des Verschuldungsgrades und Risikos (Schaubild 11). Die endgültigen Einzelziele für die Geschäftsbereiche unterscheiden sich erheblich von den ursprünglichen Einzelvorschlägen der Geschäftsbereiche, da für Kunststoffe kein Wachstum vorgesehen ist (Schaubild 12).

LITERATURANGABEN

- [1] Markowitz, H.M., Portfolio Selection, Efficient Diversification of Investments.
Second printing. New York - London - Sydney 1965
- [2] Peters, L., Simultane Produktions-Investitionsplanung mit Hilfe der Portfolio Selection.
Berlin 1971
- [3] Sharpe, W.F., A Simplified Model of Portfolio Selection.
In: MS, 9 (1963), S. 277-293

Schaubild 1

Die Zielfindung des Konzerns wird durch eine Reihe von Modellen unterstützt . . .

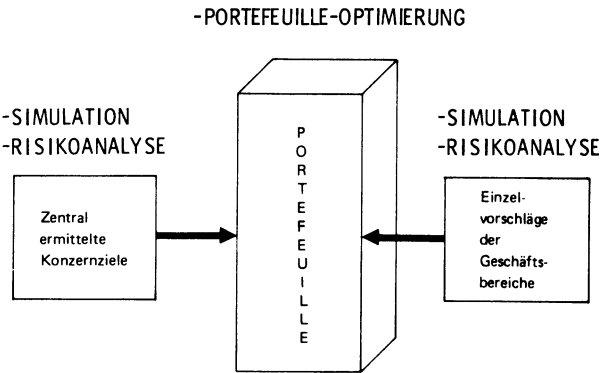


Schaubild 2

Konzernziel und Geschäftsbereichsziele müssen im Erwartungswert und Risiko aufeinander abgestimmt werden . . .

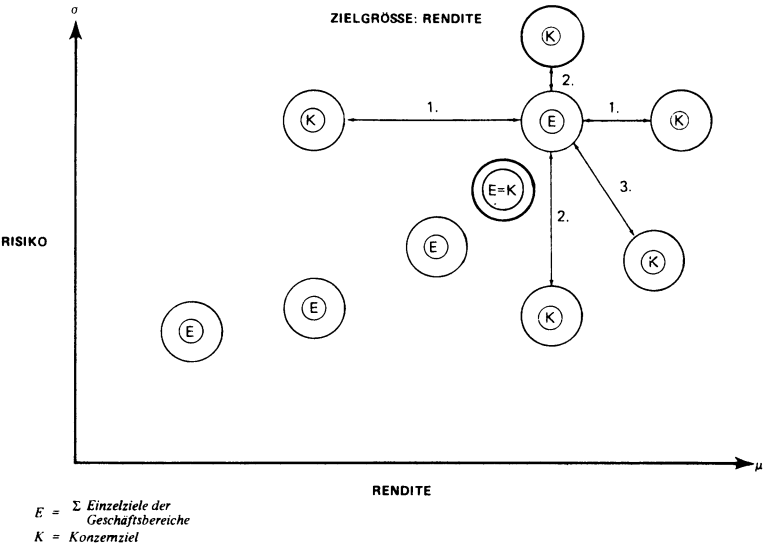


Schaubild 3

FIKTIVE ZAHLEN

Umsatzprognose

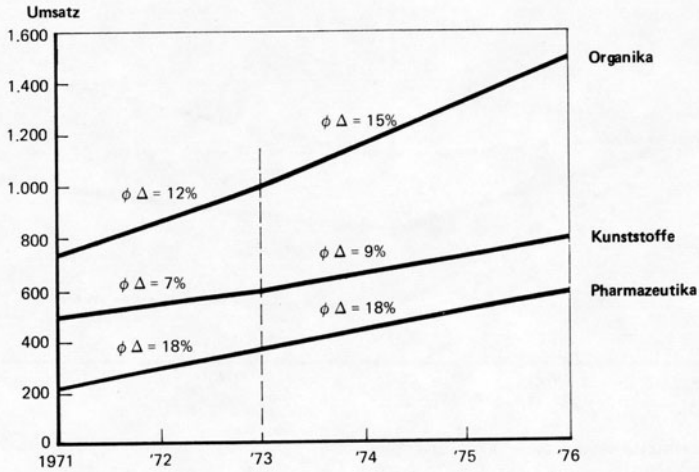


Schaubild 4

FIKTIVE ZAHLEN

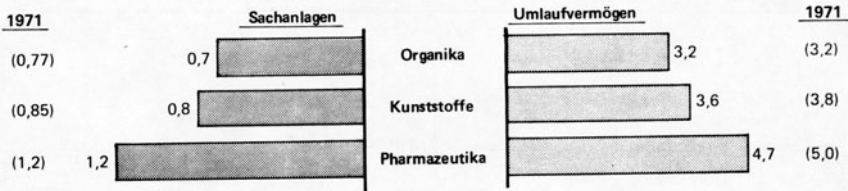
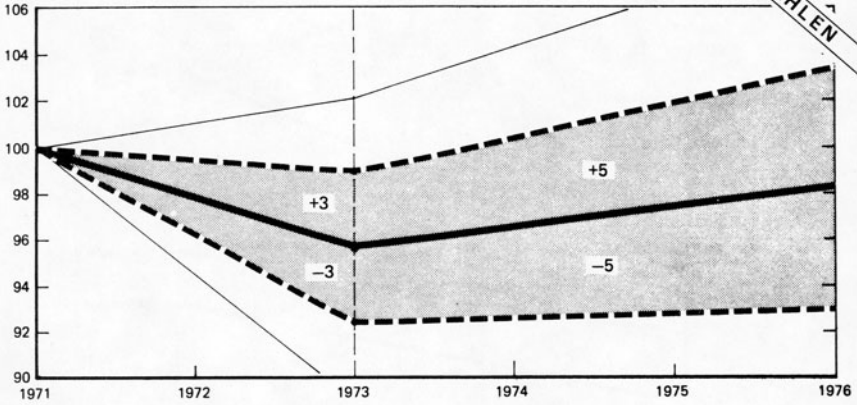
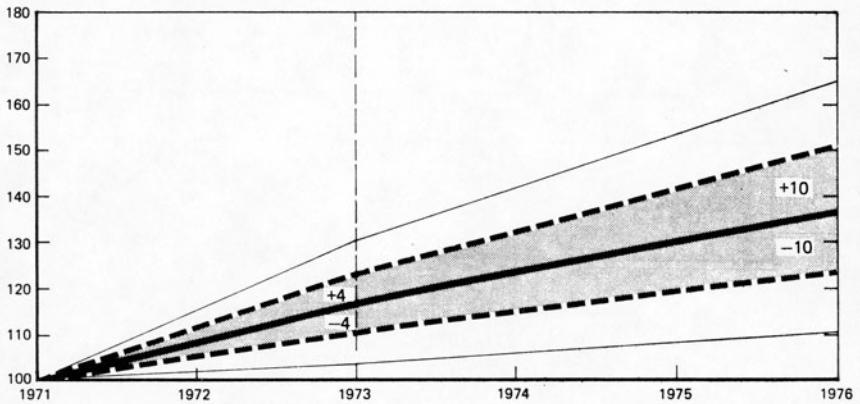
Prognose marginaler Kapitalumschlagshäufigkeiten
(ϕ 1971 - 1976)

Schaubild 5

Prognose Preisindex chemische Industrie (1971 = 100)
(fiktive Zahlen)



Prognose Kostenindex chemische Industrie (1971 = 100)
(fiktive Zahlen)



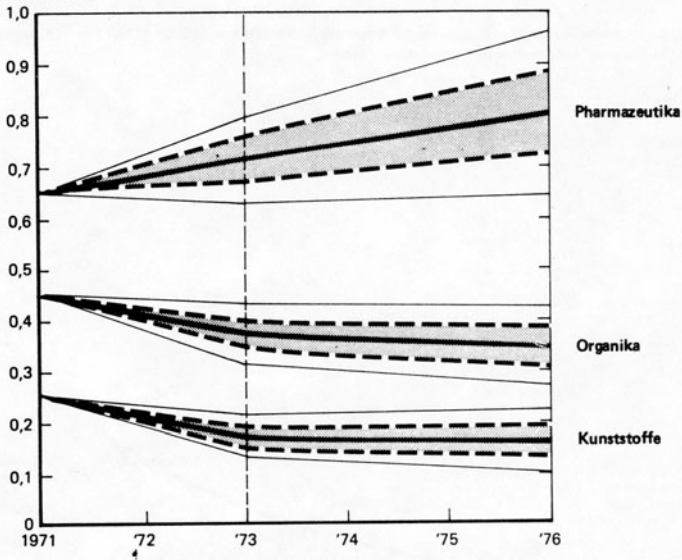
Streuungsbereich der Indizes:



Schaubild 6
FIKTIVE ZAHLEN

**Prognose der Bruttogewinnbeiträge in Abhängigkeit von Preis- und Kostenindizes
der chemischen Industrie**

Bruttogewinnbeitrag
pro 1 DM Erlös



Index (1971 = 100) Chemische Industrie

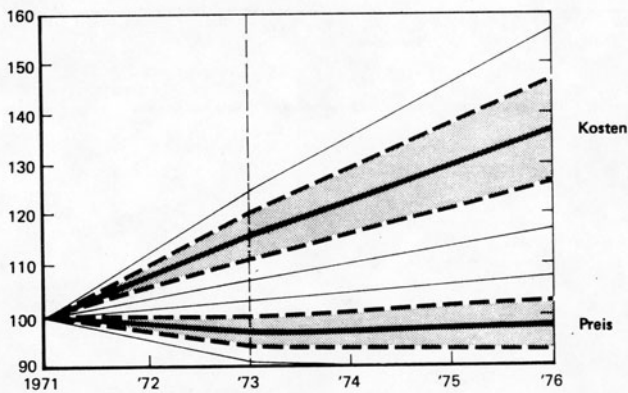
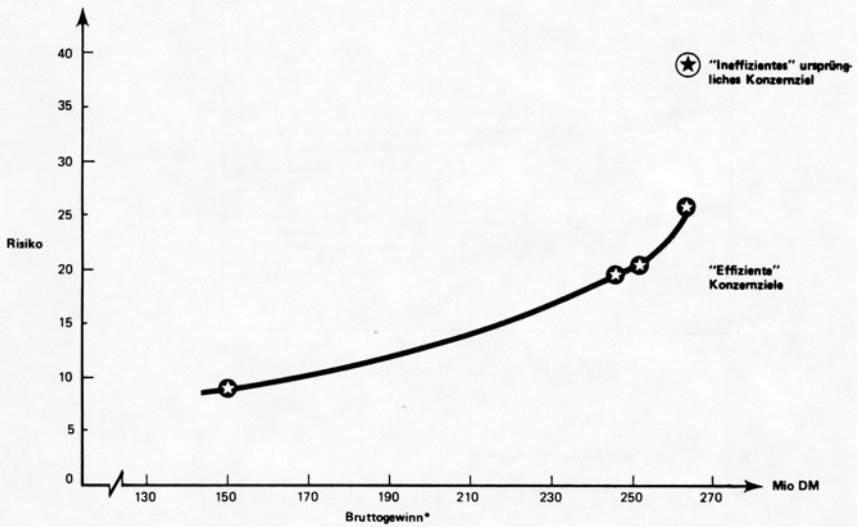


Schaubild 7
FIKTIVE ZAHLEN

Das ursprüngliche Gewinn-Konzernziel (Grundlage der Gesamtkapitalrendite von 16% für 1972-73 und 13% für 1974-76) hat sich als "ineffiziente" Gewinn-Risiko-Kombination erwiesen . . .



* Gesamtzeitraum 1972 - 1976, nach Abzug fixer Kosten von Investitionen, kalk. Kosten u. Zinsen

FIKTIVE ZAHLEN

Schaubild 8

Unter Berücksichtigung der Abhängigkeiten zwischen Geschäftsbereichen kann sich das Risiko vermindern . . .

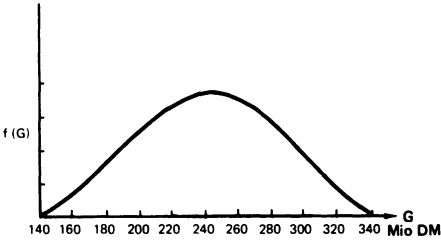
Ohne Abhängigkeiten

Risiko

Geschäftsbereiche einzeln *

	ungünstig $\mu - \sigma$	erwartet μ	günstig $\mu + \sigma$
Organika	67	87	107
Kunststoffe	-7	0	7
Pharmazeutika	135	154	173
Gesamt	197	241	285

Geschäftsbereiche einzeln

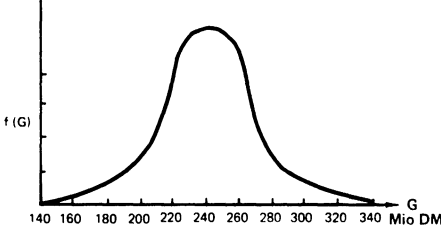


Mit Abhängigkeiten

Geschäftsbereiche Portfeuille *

Gesamt	217	241	265
--------	-----	-----	-----

Geschäftsbereiche Portfeuille



* Bruttogewinnzahlen in Mio DM für 1974-76

Die Wahl bestimmter Einzelziele ist von der Risikopräferenz der Konzernleitung abhängig . . .

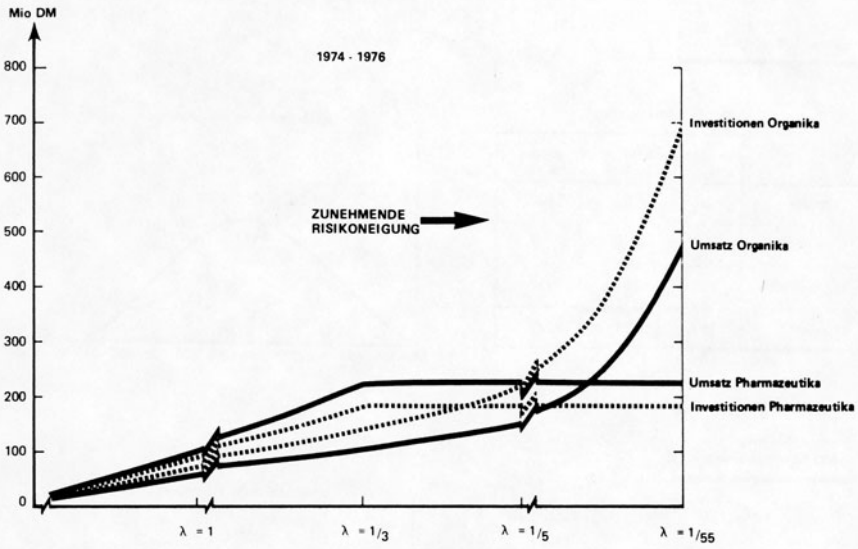
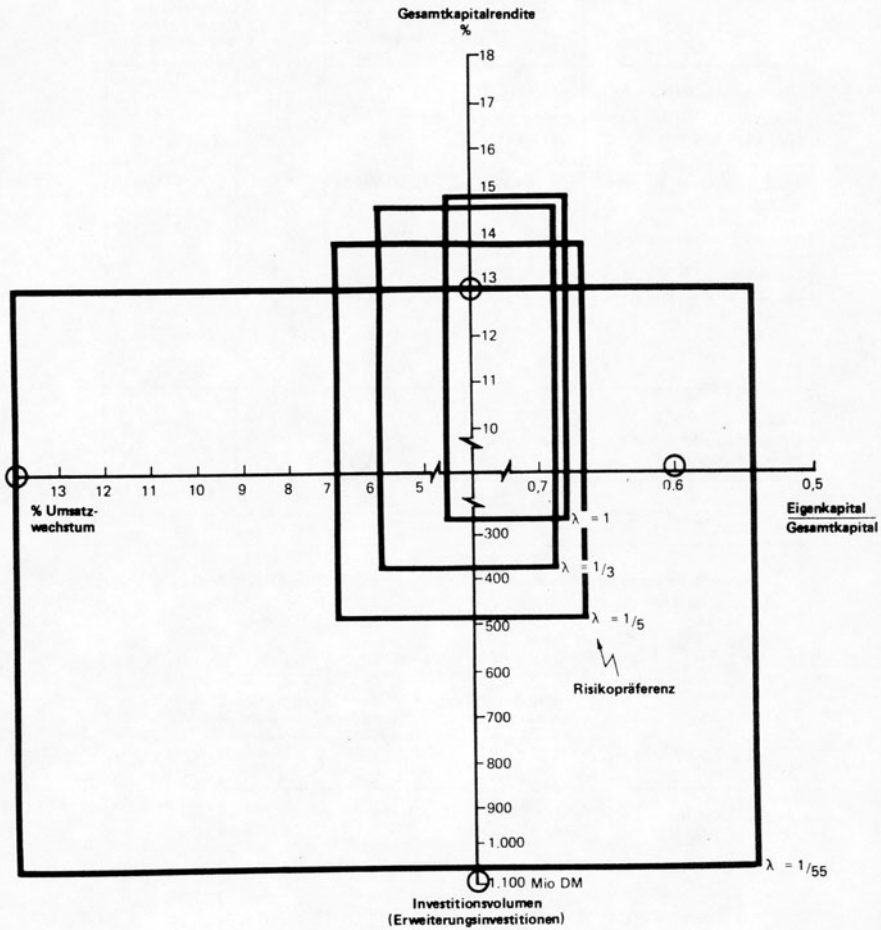


Schaubild 10

Die ursprünglichen Konzernziele 1976 lassen sich nur bei äußerster Risikopräferenz und mit einer höheren Verschuldung realisieren . . .



○ ursprüngliche Konzernziele

Schaubild 11
FIKTIVE ZAHLEN

Während sich die Konzernziele nur in der Verschuldung und im Risiko ändern . . .

	Konzernziele 1974 - 1976		Summe Einzelvorschläge der Geschäftsbereiche 1974 - 1976
	Ursprünglich	Modell (max. Risikoeignung) = endg. Ziels.	
Gesamtkapitalrendite	13%	13%	11%
ϕ Umsatzwachstum	14%	14%	16%
Zusätzl. Investitionsvolumen	1.100 Mio	1.100 Mio	1.343 Mio
Verschuldung (Eigenkapital/Gesamtkapital)	0,6	0,55	0,5
Dividende	15%	15%	15%

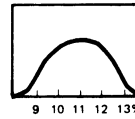
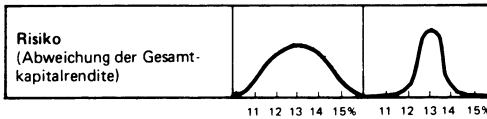


Schaubild 12
FIKTIVE ZAHLEN

. . . weichen die endgültigen Einzelziele im Geschäftsbereich Kunststoffe erheblich von den ursprünglichen Vorschlägen ab

	1974 - 1976					
	Einzelvorschläge der Geschäftsbereiche			Modellvorschlag = endgültige Zielsetzung		
	U	G	I	U	G	I
Organika	15%	87	853	15%	87	853
Kunststoffe	9%	0	258	—	—	—
Pharmazeutika	18%	154	232	18%	154	232

U = ϕ Umsatzwachstum
G = Zusatzl. Bruttogewinn (Mio DM)
I = Zusatzl. Investitionen (Mio DM)

Mehrstufige Planung einzelner Produkte und Projekte bei Unsicherheit

von M. Henke, Bonn

1. Einführung

Dieser Beitrag ist im wesentlichen eine Fortsetzung der Arbeiten HENKE [1], [2], wo einmal Probleme des Stoppens von Forschungsprojekten - etwa aus Gründen sinkender Rentabilität - behandelt wurden und zum anderen untersucht wurde, ob und wann ein Produkt, dessen Produktion infolge schrumpfender Absatzmengen bedenklich geworden ist, gestoppt werden soll. Dabei stellte es sich heraus, daß beide Arten dieser mehrstufigen, sequentiellen Entscheidungsprozesse im allgemeinen mit den gleichen Methoden behandelt werden können. Allerdings muß man die jeweils auftretenden zufälligen Periodengewinne - generell als Y_i , $i = 1, 2, \dots, N$, bezeichnet - für die speziellen ökonomischen Probleme unterschiedlich definieren. So kann man etwa fordern:

$$Y_i = I(e_i)G_i - \{1 - I(e_i)\}W_i, \quad I(e_i) = \begin{cases} 1, & \text{für } e_i \\ 0, & \text{sonst} \end{cases} \quad (1)$$

$i = 1, 2, \dots, N$

$$Y_i = (p_i - k_i)X_i - k_0(i) \quad (2)$$

Die Darstellung (1) ist sinnvoll bei Forschungs- und Entwicklungsprojekten, wobei G_i den zufälligen Periodengewinn aus dem entwickelten Projekt (Produkt) angibt, W_i die zufälligen Forschungskosten pro Periode kennzeichnet, und e_i charakterisiert das Ereignis, das Projekt spätestens in der Periode i fertigzustellen. Die Darstellung (2) ist möglich bei Produkt-Abbruch-Entscheidungen, mit: p_i = Stückpreis, k_i = Stückkosten (variabel), X_i = zufällige Absatzmenge, $k_0(i)$ = Fixkosten, jeweils in der Periode i .

Bei Produkten oder Projekten gibt es neben der einfachen Stoppentscheidung noch weitere Entscheidungsmöglichkeiten, z.B. den Einsatz zusätzlicher Werbung, eine neue Preispolitik, die Erschließung neuer Märkte, oder die Ersetzung eines alten

Produktes oder Forschungsprojektes durch ein attraktives neues. Alle diese Entscheidungsmöglichkeiten sollen im folgenden Umschaltungsmöglichkeit genannt werden von dem Prozess (a) mit den Gewinnen Y_i auf den Prozess (b) mit den Gewinnen Z_1, Z_2, \dots, Z_N , wobei die Verteilungen der Y_i und Z_i bekannt sind und die Erwartungswerte endlich. Dabei soll in diesem Beitrag nur der Fall einer einmaligen Umschaltung untersucht werden. Bei mehreren gleichzeitig zugelassenen Umschaltungen, wenn als z.B. gleichzeitig neue Werbeanstrengungen unternommen werden und neue Märkte erschlossen werden sollen, kann die Lösung in ähnlicher Weise gewonnen werden. Die Zahl der Zufallsgrößen und damit die Zahl der Zustandsvariablen steigt allerdings proportional zu der Zahl der Umschaltungen. Für spezielle ökonomische Probleme kann man die neuen Gewinne Z_i - auch Verluste, wenn Z_i negativ ist - wie folgt festlegen:

$$Z_i = (\hat{p}_i - \hat{k}_i) \hat{X}_i - \hat{k}_0(i) \quad (3)$$

$$Z_i = (\hat{p}_i - \hat{k}_i) (X_i + \Delta X_i) - k_0(i) - \Delta k_0(i) \quad (4)$$

$$Z_i = I(\hat{e}_i) \hat{G}_i - \{1 - I(\hat{e}_i)\} \hat{W}_i \quad (5)$$

Darstellung (3) kann angeben, daß das alte Produkt durch ein neues ersetzt wird mit den neuen Größen $\hat{p}_i, \hat{k}_i, \hat{X}_i, \hat{k}_0(i)$. (4) kann bedeuten, daß ein neuer Markt erschlossen wird und zusätzliche Werbung eingesetzt wird, wobei ΔX_i die zusätzliche (geschätzte) Absatzmenge und $\Delta k_0(i)$ die zusätzlichen Werbe- oder Markterschließungskosten sind. Das Beispiel (5) kann angeben, daß ein altes Projekt durch ein neues ersetzt wird oder das alte Projekt beschleunigt oder verlangsamt wird. Diese wenigen Beispiele zeigen, daß sich für die nachfolgenden Ergebnisse eine Reihe von ökonomischen Interpretationsmöglichkeiten ergeben.

2. Endlichstufige Prozesse bei einmaliger Umschaltung

Angenommen, der Entscheidende habe nicht nur die Wahl vor jeder Periode $i, i=1, \dots, N$, das Projekt (das Produkt) abzubauen ($a_i=0$) oder fortzusetzen ($a_i=1$), sondern könne z.B. das Projekt verlangsamen oder durch ein neues ersetzen. Er kann wählen zwischen dem Prozess (a) mit den Gewinnen Y_i und

und dem Prozess (b) mit den Gewinnen Z_i . Die Z_i und Y_i seien unabhängig und die Verteilungen der Gewinne sind bekannt.

Vorausgesetzt sei, der Entscheidende beginne im Prozess (a), dann soll er die folgenden Entscheidungsmöglichkeiten haben:

$a_i=1$ setze das Projekt (Produkt) in der Periode i im Prozess (a) fort,

$a_i=0, b_i=1$ schalte in der Periode i vom Prozess (a) auf den Prozess (b),

$a_i=0, b_i=0$ stoppe den Prozess vor der i -ten Periode.

Wählt der Entscheidende $a_i=0, b_i=1$, so hat von der $(i+1)$ -ten Periode an, nur noch die Wahl $b_k=1$: setze den Prozess (b) fort oder $b_k=0$: stoppe (auch) den Prozess (b), $k=i+1, i+2, \dots, N$.

Die Weiteren Voraussetzungen entsprechen im wesentlichen denen des Grundmodells in Henke [2], so ist zu Beginn des Gesamtprozesses ein Kapitalbestand von j Einheiten vorhanden und $U_1(j) < \infty$ ist der kardinale meßbare Nutzen von j Kapitaleinheiten, wenn der Prozess vor der Periode i gestoppt wird. Für den Gegenwartsnutzen muß noch die Diskontierung mit β , $0 \leq \beta \leq 1$, berücksichtigt werden. Das Kapital, das sich aus dem Prozess ergibt, wird produktiv im Unternehmen eingesetzt und vermehrt sich so pro Periode mit dem Faktor r , $r \geq 1$, oder es wird bei Bedarf Kapital aus dem Unternehmen für den Prozess zur Verfügung gestellt. Eine Kapitalrationierung ist nicht vorgesehen. (Eine Erweiterung ist allerdings entsprechend zu Henke [2, Kap. 4] sofort möglich).

Stellt man beispielhaft für 3 Perioden den jeweiligen Entscheidungen die entsprechenden Nutzen (Kapitalwerte) gegenüber, so zeigt sich die Konstruktion des Modell deutlich:

Entscheidungen	"zufällige" Kapitalwerte
$a_1=0, b_1=0$	$U_1(j)$
$a_1=0, b_1=1, b_2=0$	$\beta U_2((j+Z_1)r)$
$a_1=0, b_1=1, b_2=1, b_3=0$	$\beta^2 U_3((j+Z_1)r+Z_2)r$
$a_1=1, a_2=0, b_2=0$	$\beta U_2((j+Y_1)r)$
$a_1=1, a_2=0, b_2=1, b_3=0$	$\beta^2 U_3((j+Y_1)r+Z_2)r$
$a_1=1, a_2=1, a_3=0, b_3=0$	$\beta^2 U_3((j+Y_1)r+Y_2)r$

Die folgende schematische Darstellung, wo $j \leftrightarrow Y_1$ bedeutet, daß das Kapital um Y_1 zunehmen kann, verdeutlicht für den allgemeinen Fall N die Asymmetrie des Problems, da man zwar von Prozess (a) zum Prozess (b), aber nicht umgekehrt von (b) nach (a) schalten kann:

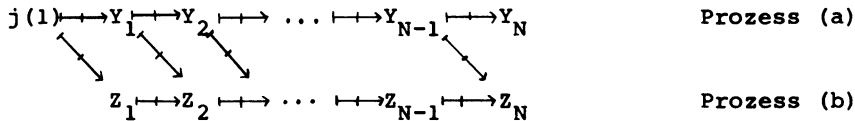


Abb. 1: Einfacher Umschaltprozess

Das Symbol \leftrightarrow wurde dabei verwendet, um anzugeben daß das Projekt vor jedem $i=1, \dots, N$ auch abgebrochen werden kann. Für das Beispiel $N=2$ ergibt sich für die Maximierung des erwarteten Kapitalwertes ohne Kapitalrationierung:

$$\begin{aligned}
 Q_j^{(1)} = \max_{B_1} E & \left[(1-a_1)(1-b_1)U_1(j) + \beta \{ b_1(1-a_1)(1-b_2)U_2([j+Z_1]r) \right. \\
 & + a_1(1-a_2)(1-b_2)U_2([j+Y_1]r) + \beta b_1 b_2(1-a_1)(1-b_3) \\
 & U_3([(j+Z_1)r+Z_2]r) \\
 & + \beta a_1 b_2(1-a_2)(1-b_3)U_3([(j+Y_1)r+Z_2]r) + \beta a_1 a_2(1-a_3) \\
 & \left. (1-b_3)U_3([(j+Y_1)r+Y_2]r) \} \right] \quad (6)
 \end{aligned}$$

mit

$$B_1 = \{a_1, b_1, a_2, b_2, a_3, b_3 \mid a_k, b_k \in \{0, 1\}, a_k + b_k \leq 1, k=1, 2; a_3, b_3 = 0\}$$

(Das Maximum in (6) soll existieren; Existenz

probleme entstehen in Kap. 3) (7)

Kürzt man den Ausdruck in $\{ \}$ aus (6) ab mit R und definiert:

$R = H(a_1, b_1, a_2, b_2, a_3, b_3)$, so erhält man:

$$Q_j^{(1)} = \max_B E \{ (1-a_1)(1-b_1)U_1(j) + \beta H(a_1, b_1, a_2^*, b_2^*, a_3^*, b_3^*) \} \quad (8)$$

mit:

$$B = \{a_1, b_1 \mid a_1, b_1 \in \{0, 1\}, a_1 + b_1 \leq 1\}, \quad (9)$$

dabei sind $a_k^*, b_k^*, k=2,3$ die optimal bestimmten a_k, b_k . Vereinfacht man weiter:

$$E(H(a_1, b_1, a_2^*, b_2^*, a_3^*, b_3^*)) = G(a_1, b_1) \quad (10)$$

und setzt voraus, daß die Entscheidung bei $i=1$ erst getroffen werden muß, wenn die Gewinne Y_1, Z_1 der ersten Periode bekannt sind, also $Y_1=y, Z_1=z$, so gilt mit der Vereinfachung

$$Q_j^{(1)} | Y_1=y, Z_1=z := Q_j^{(1)} | y, z$$

$$Q_j^{(1)} | y, z = \text{MAX} \left[\begin{array}{ll} U_1(j) & , \text{ falls } a_1=0, b_1=0 \\ \beta \cdot G(1, b_1) | y, z & , \text{ falls } a_1=1 \\ \beta \cdot G(0, 1) | y, z & , \text{ falls } a_1=0, b_1=1 \end{array} \right] \quad (11)$$

Weiter ist aus (6)-(10):

$$\begin{aligned} G(1, b_1) | y, z = \text{MAX}_{B_2} E \{ & (1-a_2)(1-b_2)U_2(\{(j+y)r\}) + \beta b_2(1-a_2)(1-b_3) \cdot \\ & U_3(\{(j+y)r+Z_2\}r) + \beta a_2(1-a_3)(1-b_3) \\ & U_3(\{(j+y)r+Y_2\}r) \} := Q_{(j+y)r}^{(2)} \end{aligned} \quad (12a)$$

mit:

$$B_2 = \{a_2, b_2, a_3, b_3 \mid a_2, b_2 \in \{0, 1\}, a_2 + b_2 \leq 1; a_3, b_3 = 0\}. \quad (12b)$$

Die Definition $Q_{(j+y)r}^{(2)}$ ist berechtigt, da $Q^{(1)}$ definiert ist wie $Q^{(2)}$, wenn man den Index um "eins" erhöht und an die Stelle von j : $(j+y)r$ setzt und beachtet, daß der Prozeß nach $N=2$ beendet ist.

Für $G(0, 1) | y, z$ folgt:

$$\begin{aligned} G(0, 1) | y, z = \text{MAX}_{B_2} E \{ & (1-b_2)U_2(\{(j+z)r\}) + \beta b_2(1-b_3) \\ & U_3(\{(j+z)r+Z_2\}r) \} := F_{(j+z)r}^{(2)}. \end{aligned} \quad (13a)$$

$$\text{mit: } B_2 = \{b_2, b_3 \mid b_2 \in \{0, 1\}, b_3 = 0\}, \quad (13b)$$

wobei $F_j^{(i)}$ wie in Henke [2, Res.1] definiert ist, s. auch (15b). Die Definition $F_{(j+z)r}^{(2)}$ ist berechtigt, da $G(0,1) \mid y, z$ eben genauso definiert ist wie $F_j^{(i)}$ für den Spezialfall $i=2$ und $j(2) = \{j + z\}r$.

Damit erhält man aus (11), (12), (13):

$$Q_j^{(1)} \mid y, z = \text{MAX}(U_1(j), BQ_{(j+y)r}^{(2)}, BF_{(j+z)r}^{(2)}). \quad (14)$$

Läßt man wieder allgemein zu, daß $\beta = \beta(i), r = r(i)$ und setzt für $1:i$ für $2:i+1$ so gilt für den Prozesswert unter der Bedingung, daß bereits $i-1$ Perioden vorüber sind und $j = j(1)$ der Kapitalbestand ist:

Resultat 1:

(a) Für den Prozesswert im Typ (F_1) folgt:

$$Q_j^{(i)} = E\{\text{MAX}(\beta(i)Q_{(j+Y_1)r(i)}^{(i+1)}, \beta(i)F_{(j+Z_1)r(i)}^{(i+1)}, U_1(j))\}, \quad (15a)$$

mit:

$$F_j^{(i)} = E\{\text{MAX}(\beta(i)F_{(j+Z_1)r(i)}^{(i+1)}, U_1(j))\}, \quad i=1, 2, \dots, N, \quad \text{alle } j \quad (15b)$$

$$Q_j^{(N+1)} = F_j^{(N+1)} = U_{N+1}(j), \quad (\text{Randbedingung}). \quad (15c)$$

(b) Für die optimale Wahl- bzw. Stoppregel ergibt sich:

(b₁) falls das Projekt vor der Periode i im Prozess (a) behandelt wird: Stoppe das Projekt beim ersten i

$$a_i^* = 0, \quad b_i^* = 0, \quad \text{wenn } U_1(j) \geq \beta(i) \text{MAX}(Q_{(j+Y_1)r(i)}^{(i+1)}, F_{(j+Z_1)r(i)}^{(i+1)}), \quad (16a)$$

setze das Projekt in der Periode i fort im Prozess (a)

$$a_i^* = 1, \quad \text{wenn } Q_{(j+Y_1)r(i)}^{(i+1)} \geq F_{(j+Z_1)r(i)}^{(i+1)} \quad (16b)$$

$$\text{und } \beta(i)Q_{(j+Y_1)r(i)}^{(i+1)} > U_1(j),$$

schalte von (a) nach (b) in der Periode i

$$a_i^*=0, b_i^*=1, \text{ wenn } F_{(j+Z_i)}^{(i+1)} r(i) > \text{MAX}(U_i(j), \beta Q_{(j+Y_i)}^{(i+1)} r(i)), \quad (16c)$$

(b₂) falls das Projekt vor k im Prozess (b) behandelt wird:
Stoppe das Projekt beim ersten k dann und nur dann

$$b_k^*=0, \text{ wenn } U_k(j) \geq \beta(k) F_{(j+Z_k)}^{(k+1)} r(k), \quad k=i+1, \dots, N \quad (17)$$

- - - - -

Zum Beweis sei angeführt, daß sich der Prozesswert in (15a) sofort aus den vorausgegangenen Umformungen (6) - (14) ergibt, wenn man das Beispiel N=2 auf den allgemeinen Fall N überträgt.

Die optimale Politik für den Fall (b₁), ergibt sich nun unmittelbar aus (11) - (14), da z.B. gemäß (11) dem Fall a₁=0, b₁=0 die Auszahlung U₁(j) zugeordnet ist und für die optimale Politik jeweils das Maximum aus den drei Möglichkeiten gewählt werden muß. Der Fall der Gleichheit zwischen Prozess (a) und Prozess (b), also:

$Q_{(j+Y_i)}^{(i+1)} r(i) = F_{(j+Z_i)}^{(i+1)} r(i) > \frac{1}{\beta(i)} U_i(j)$ wurde der Entscheidung: setze den Prozess (a) fort zugeordnet, aus der Überlegung, daß es im allgemeinen bequemer ist, den Prozess fortzusetzen als umzuschalten. (Die Zuordnung ist nicht zwingend, da sich der gleiche (optimale) Prozesswert auch dann ergibt, wenn für diesen Fall von (a) nach (b) umgeschaltet wird),

Für den Fall (b₂), wenn bereits auf den Prozess (b) umgeschaltet wurde, besteht nur noch die Wahl b_k=0 oder b_k=1. Die optimale Politik ergibt sich dabei entsprechend.

Für die Definition der optimalen Stoppvariablen ist es wesentlich, daß das Projekt entweder abgebrochen wird

Die doppelten Pfeile geben z.B. an, daß für den Kapitalbestand am Ende der 4-ten Periode oder vor der 5-ten gilt, falls das Projekt nicht abgebrochen wird:

$$j(5) = j(1) + y_1 + z_2^{(2)} + z_3^{(2)} + z_4^{(2)},$$

insbesondere wurde also hier bei $i=2$ auf den Prozess $(b)_2$ umgeschaltet.

Ohne nochmals auf die formalen Einzelheiten einzugehen, vgl. (6) - (14), sei angemerkt, daß $F_{(j+Z_i)}^{(i+1)} r(i)$ in Res. 1 durch $F_{(j+Z_i)}^{(i+1)} r(j)$ ersetzt werden muß und somit gilt:

Resultat 2:

(a) Der Prozesswert im Prozesstyp (F_2) ist:

$$\hat{Q}_j^{(i)} = E\left\{ \max\left(\beta(i) \hat{Q}_{(j+Y_i)}^{(i+1)} r(i), \beta(i) \hat{F}_{(j+Z_i)}^{(i+1)} r(i), U_i(j) \right) \right\}, \quad i=1, \dots, N, \text{ alle } j \quad (19a)$$

mit:

$$\hat{F}_j^{(k)} = E\left\{ \max\left(\beta(k) \hat{F}_{(j+Z_k)}^{(k+1)} r(k), U_k(j) \right) \right\}, \quad k=i+1, \dots, N \quad (19b)$$

und:

$$\hat{Q}_j^{(N+1)} = \hat{F}_j^{(N+1)} = U_{N+1}(j) \quad (19c)$$

(b) Die optimale Politik ist gegeben wie im Res. 1 (b) wenn jeweils $Q_j^{(i)}, F_j^{(i)}$ aus (15) durch die Werte $\hat{Q}_j^{(i)}, \hat{F}_j^{(i)}$ ersetzt werden.

Ebenso wie in den einfacheren Prozessen, s. Henke [1], [2], können wieder die Nettoprozesswerte bestimmt werden. So gilt für Prozesstyp (F_2) :

$$\hat{q}_j^{(i)} := \hat{Q}_j^{(i)} - U_i(j); \quad \hat{f}_j^{(i)} := \hat{F}_j^{(i)} - U_i(j) \quad (20)$$

Es soll aber darüberhinaus sofort der Spezialfall $\beta(i) \cdot r(i)=1, U_i(j)=j$ vorausgesetzt werden, da sich sonst keine

formale Vereinfachung ergibt. Definiert man zur Abkürzung $E(\text{MAX}(Z, 0)) := E(Z)^+$, Z bel. Zufallsgröße, so folgt:

Resultat 3:

(a₁) Der Nettoprozesswert im Prozesstyp (F₁) ist für den Fall $r(i)\beta(i)=1$, alle i und $U_1(j)=j$:

$$q(i) = E\{\text{MAX}(Y_1 + \beta(i)q(i+1), Z_1 + \beta(i)f(i+1), 0)\} \quad (21a)$$

$$f(i) = E\{Z_1 + \beta(i)f(i+1)\}^+ \quad (21b)$$

$$q(N+1) = f(N+1) = 0 \quad (21c)$$

(a₂) Der Nettoprozesswert im Prozesstyp (F₂) für den Fall $\beta(i)r(i)=1$ und $U_1(j)=j$ ist

$$\hat{q}(i) = E \text{ MAX}(Y_1 + \beta(i)\hat{q}(i+1), Z_1^{(i)} + \beta(i)\hat{f}(i+1), 0), i=1, 2, \dots, N \quad (22a)$$

$$\hat{f}(k) = E\{Z_k^{(i)} + \beta(k)\hat{f}(k+1)\}^+, k=i+1, \dots, N \quad (22b)$$

$$\hat{q}(N+1) = \hat{f}(N+1) = 0, \quad (22c)$$

insbesondere sind also die Nettoprozesswerte (wenn $i-1$ Perioden bereits vorüber sind) unabhängig vom Kapitalbestand $j=j(i)$.

Beweis:

Der Beweis sei geführt für den Prozess (F₂). Wie man sieht ist

$$\hat{f}_j^{(k)} = \hat{F}_j^{(k)} - U_k(j) \text{ unabhängig von } j, \text{ für alle } k=i+1, \dots, N.$$

So gilt für $k=N$ und ein bel. aber festes $i \leq k-1$ aus (19b)

$$\hat{f}_j^{(N)} = \hat{F}_j^{(N)} - U_N(j) = E\{\text{MAX}(\beta(N)\{j + Z_N^{(i)}\}r(N) - j, 0)\}$$

$$= E\{Z_N^{(i)}\}^+ := \hat{f}(N) \text{ für jedes feste } i \in \{1, \dots, N-1\}.$$

Vorausgesetzt $\hat{f}_j^{(k+1)}$ ist unabhängig von j für bel. aber festes $i, i \leq k-1$, dann gilt:

$$\hat{f}_j^{(k)} = E \left\{ \text{MAX}(\beta(k) \{ \hat{F}^{(k+1)}_{\{j+Z_k^{(i)}\}} r(k) \} - \{j+Z_k^{(i)}\} r(k) \} + j+Z_k^{(i)} - j, 0) \right\}$$

$$= E \{ \text{MAX}(Z_k^{(i)} + \beta(k) \hat{f}^{(k+1)}, 0) \} := \hat{f}^{(k)} \quad \text{unabhängig}$$

von j .

Da nun $\hat{f}_j^{(k)}$ unabhängig ist von j für alle $k=i+1, \dots, N$ gilt zunächst aus (19a)

$$\begin{aligned} \hat{q}_j^{(i)} &= E \left\{ \text{MAX}(\beta(i) \cdot \hat{Q}^{(i+1)}_{(j+Y_1)r(i)} - j, \beta(i) \{ \hat{F}^{(i+1)}_{\{j+Z_1^{(i)}\}} r(i) \} \right. \\ &\quad \left. - (j+Z_1^{(i)}) r(i) \} + Z_1^{(i)}, 0) \right\} \\ &= E \text{MAX}(\beta(i) \hat{Q}^{(i+1)}_{\{j+Y_1\}r(i)} - j, Z_1^{(i)} + \beta(i) \hat{f}^{(i+1)}, 0). \end{aligned} \quad (24)$$

Setzt man wie vorher als Induktionsvoraussetzung, daß

$$\hat{q}_j^{(i+1)} := \hat{q}^{(i+1)}, \text{ alle } j, \text{ folgt aus (24)}$$

$$\begin{aligned} \hat{q}_j^{(i)} &= E \left[\text{MAX}(\beta(i) \{ \hat{Q}^{(i+1)}_{(j+Y_1)r(i)} - (j+Y_1)r(i) \} + Y_1 \right. \\ &\quad \left. , Z_1^{(i)} + \beta(i) \hat{f}^{(i+1)}, 0) \right], \end{aligned}$$

wegen:

$$\hat{Q}^{(i+1)}_{\{j+Y_1\}r(i)} - (j+Y_1)r(i) = \hat{q}^{(i+1)}_{(j+Y_1)r(i)} := \hat{q}^{(i+1)} \quad \text{ergibt}$$

sich das Ergebnis (22a) sofort.

Das Res. 3 ergibt somit eine wesentliche formale Vereinfachung, da man $\hat{q}^{(i)}$, abhängig nur von der jeweiligen Entscheidungsstufe $i=1, \dots, N$, einfacher errechnen kann als $\hat{Q}_j^{(i)}$ das von zwei Parametern (i, j) bestimmt wird. Es ist einsichtig, daß man die Prozesstypen auch auf den Fall von Markovprozessen ausdehnen kann. Dazu gehen wir aus vom Prozesstyp F_1 und nehmen über die Verteilungen der Y_1, Z_1 an:

$$P\{Y_1=y\} = p_1(y) ; P\{Z_1=z_1\} = p_1(z_1)$$

$$P\{Y_i=y_i | Y_{i-1}=y_{i-1}\} = p_1(y_i | y_{i-1}) , i \geq 1, y_i, y_{i-1} \in I_1$$

$$P\{Z_k=z_k | Z_{k-1}=z_{k-1}\} = p_k(z_k | z_{k-1}) , k \geq i+1, z_k, z_{k-1} \in I_2,$$

I_1, I_2 sind endliche Zustandsmengen.

Da in Markovprozessen die Prozesswerte von den vorausgegangenen Zuständen abhängen, setzen wir für $q(i): v_i(y_{i-1})$ und für $f(k): m_k(z_{k-1})$ und erhalten so:

Resultat 4:

(a) Der Nettoprozesswert im Markovprozess (F_3) ist:

$$v_i(y_{i-1}) = \sum_{y_i \in I_1} \sum_{z_i \in I_2} \text{MAX}(y_i + \beta(i) v_{i+1}(y_i), z_i + \beta(i) m_{i+1}(z_i), 0) p_1(y_i | y_{i-1}) p_1(z_i), i=2, \dots, N, \quad (24a)$$

mit:

$$m_k(z_{k-1}) = \sum_{z_k \in I_2} \text{MAX}(z_k + \beta(k) m_{k+1}(z_k), 0) p_k(z_k | z_{k-1}) \quad (24b)$$

$$\{v_{N+1}(y_N), m_{N+1}(z_N)\} := \{0, 0\} \quad (\text{Randbedingung}) \quad (24c)$$

(b) Die optimale Politik ergibt sich aus dem Vergleich der Maximanden in (24a, b) als:

$$a_i^* = 0, b_i^* = 0, \text{ wenn } 0 \geq \text{MAX}(y_i + \beta(i) v_{i+1}(y_i), z_i + \beta(i) m_{i+1}(z_i)),$$

$$a_i^* = 1, \text{ wenn } y_i + \beta(i) v_{i+1}(y_i) \geq z_i + \beta(i) m_{i+1}(z_i) \text{ und}$$

$$y_i + \beta(i) v_{i+1}(y_i) > 0,$$

$$a_i^* = 0, b_i^* = 1, \text{ wenn } z_i + \beta(i) m_{i+1}(z_i) > \text{MAX}(y_i + \beta(i) v_{i+1}(y_i), 0),$$

$$b_k^* = 0 \iff z_k \leq -\beta(k) m_{k+1}(z_k).$$

3. Unendlichstufige Prozesse bei einmaliger Umschaltung

Obwohl aus praktischer Sicht Grenzwertbetrachtungen für einen unendlichen Horizont insbesondere wegen der Voraussetzung, daß alle Daten (z.B. die Verteilungen) von der Periode i unabhängig sein müssen, relativ uninteressant sind, sei dennoch wegen der theoretischen Bedeutung eine solche Betrachtung vorgenommen, da man zu einfacheren und analytischen Ergebnissen gelangt.

Wir gehen aus von Res. 3 (a_1) und nehmen - wie immer bei dieser Art von Grenzbetrachtungen - eine Umindizierung vor. An die Stelle von $q(1)$ setzen wir $q(n)$, an die Stelle von $f(1):f(n)$, wobei $n=N-i+1$ die Zahl der noch folgenden Perioden angibt. Dann ergibt sich aus (21), wenn $\beta(i)=\beta$ ist und jeweils die Y_1 bzw. Z_1 identisch verteilt sind ($Y_1=Y, Z_1=Z$):

$$q(n) = E\{\text{MAX}(Y+\beta q(n-1), Z+\beta f(n-1), 0)\} \quad (25a)$$

$$f(n) = E\{Z+\beta f(n-1)\}^+ \quad (25b)$$

$$q(0) = f(0) = 0 \quad (25c)$$

Resultat 5:

(a) Für die Nettoprozesswerte in Typ (F_{1a}) gilt aus (25)

$\lim_{n \rightarrow \infty} q(n) = q^* < \infty$ und $\lim_{n \rightarrow \infty} f(n) = f^* < \infty$ existieren:

(a₁) immer für $\beta < 1$;

(a₂) für $\beta=1$ dann, wenn

$$E(Z), E(Y) \leq 0 \text{ und } |Y|, |Z| \leq C < \infty$$

Ist die Bedingung $|Y|, |Z| \leq C < \infty$ nicht erfüllt, sind $E(Y), E(Z) \leq 0$ nur notwendige Bedingungen. Gilt nur $|Y| \leq C < \infty$ ist hinreichend $E(Y) \leq 0, E(Z) \leq 0$, gilt dagegen $|Z| \leq C < \infty$ ist hin-

reichend $E(Y) \leq 0, E(Z) \leq 0$. Existiert weder für Y noch Z eine obere Schranke, so ist $E(Y), E(Z) < 0$ hinreichende Bedingung.

(b) Falls die Grenzwerte $q(n), n \rightarrow \infty (q^*)$ und $f(n), n \rightarrow \infty (f^*)$ existieren, vgl. $(a_1), (a_2)$, ergeben sie sich als:

$$q^*(1-\beta) = E\{\max(Y, Z - \beta(q^* - f^*), -\beta q^*)\}, \quad (26a)$$

$$f^*(1-\beta) = E\{\max(Z, -\beta f^*)\}, \quad (26b)$$

zusätzlich sind q^* und f^* die kleinsten Werte für die (26a,b) erfüllt ist.

(c) Die optimale Politik lautet, falls man sich im Prozess

(a) befindet und $\beta < 1$ oder $\beta = 1, E(Y), E(Z) < 0$ oder $\beta = 1, E(Y), E(Z) \leq 0, Y, Z$ diskret

$$a_1^* = 0, b_1^* = 0, \text{ wenn } -\beta q^* \geq \max(Y, Z - \beta(q^* - f^*)) \quad (27a)$$

$$a_1^* = 1, \quad \text{wenn } Y \geq Z - \beta(q^* - f^*) \text{ und } Y > -\beta q^*, \quad (27b)$$

$$a_1^* = 0, b_1^* = 1, \text{ wenn } Z > \max(Y, -\beta q^*) + \beta(q^* - f^*) \quad (27c)$$

falls bereits auf den Prozess (b) umgeschaltet wurde bei $\beta < 1$ oder $\beta = 1, E(Z) < 0$ oder $\beta = 1, E(Z) \leq 0, Z$ diskret

$$b_k^* = 0 \iff Z \leq -\beta f^*. \quad (27d)$$

Beweis:

(a₁) Da die Existenz von $f(n), n \rightarrow \infty (f^*)$ Voraussetzung ist für die Existenz von $q(n), n \rightarrow \infty (q^*)$ muß zunächst geprüft werden, wann $f^* < \infty$ existiert. Dabei weiß man aus Henke [2] daß $f^* < \infty$ immer existiert, wenn $\beta < 1$. Für $\beta = 1$ existiert $f^* < \infty$ dann, wenn $E(Z) \leq 0$ und $|Z| \leq C < \infty$. Gibt es für Z keine obere Schranke, dann ist $E(Z) \leq 0$ nur notwendig, $E(Z) < 0$ dagegen hinreichend für die Existenz von $f^* < \infty$. Damit sind also die Voraussetzungen über $Z, E(Z)$ bereits bestimmt und der Grenzwert von $f(n), n \rightarrow \infty (f^*)$ ergibt sich wie in (26b)

angegeben. Die Existenz $f(n), n \rightarrow \infty$ ist zwar notwendig, aber nicht hinreichend für $q^* < \infty$, deshalb zeigt man zunächst durch Induktion:

$$q(n) - q(n-1) \stackrel{>}{=} 0 \quad (28a)$$

$, n=1, 2, \dots < \infty$

$$q(n) - f(n) \stackrel{\geq}{=} 0 \quad (28b)$$

Kann man weiter zeigen, daß

$$\sup_n q(n) \stackrel{<}{=} \bar{C} < \infty \text{ ist, so ist die Existenz von}$$

$q(n), n \rightarrow \infty$ gesichert.

Nun ist

$$\sup_n q(n) \stackrel{<}{=} q^*, \text{ falls } q^* < \infty, q^* \text{ aus (26a)} \quad (29)$$

Da sich ansonsten, wenn für ein n $q(n-1) > q^*$ wäre, ein Widerspruch zu (28a) ergäbe. Nun ist für den Fall $\beta < 1$

$$q^* = \frac{1}{1-\beta} E\{\max(Y, Z - \beta(q^* - f^*), -\beta q^*)\} \rightarrow$$

$$q^* \stackrel{<}{=} \frac{1}{1-\beta} E\{\max(\max(Y, Z), 0)\} < \infty \quad (30)$$

da $q^* \geq 0$ ist und wegen (28b) $q^* - f^* \geq 0$. Also ist für $\beta < 1$

$\sup_n q(n) \stackrel{<}{=} q^* < \infty$ und es existiert also der Grenzwert

$q(n), n \rightarrow \infty$ immer, wenn $\beta < 1$.

(a₂) Fall: $\beta=1$. Es sind $E(Z), E(Y) \stackrel{<}{=} 0$ notwendige Bedingungen für die Existenz von $q(n), n \rightarrow \infty$, denn zunächst ist $E(Z) \stackrel{<}{=} 0$ notwendig für $f^* < \infty$. Nimmt man weiter an, daß $E(Y) > 0$ ist, braucht man nur den Prozess (a) immer weiter fortzusetzen und erhält dann:

$$q(n) = E(Y_1) + \dots + E(Y_n) = n E(Y) \text{ und also}$$

$\lim_{n \rightarrow \infty} q(n) = \infty$, deshalb muß auch $E(Y) \stackrel{<}{=} C$ sein.

Die Bedingungen $E(Y), E(Z)=0$ sind auch hinreichend, wenn $|Y|, |Z| \leq C < \infty$. Denn gibt es ein $n \geq \bar{n}$ u. $q(\bar{n}-1) \geq 2C+f^*$, gilt aus (25a) für $\beta=1$:

$$\begin{aligned} q(\bar{n}) - q(\bar{n}-1) &= E\{\text{MAX}(Y, Z - q(\bar{n}-1) + f(\bar{n}-1), -q(\bar{n}-1))\} \\ &= E(Y) \leq 0 \end{aligned}$$

Für den Fall $E(Y) < 0$ ist dies ein Widerspruch zu (28a), also muß $\sup_n q(n) \leq 2C+f^* = \bar{C} < \infty$ sein. Ist $E(Y)=0$ gilt für alle $n \geq \bar{n}$: $|q(n) - q(n-1)| = 0 \leq \delta$. Also sind in beiden Fällen die Konvergenzkriterien erfüllt.

Für den Fall, daß $|Z|, |Y| \leq C < \infty$ nicht erfüllt ist, sei zum Beweis der Fall herausgegriffen, wo nur $|Z| \leq C < \infty$. In diesem Fall ist aus Henke [2] bekannt, daß der Grenzwert $f(n), n \rightarrow \infty$ f^* existiert. Weiter steigt $q(n)$ aus (25a) monoton mit n . Außerdem kann man zeigen, daß $\sup_n q(n) \leq q^*$ ist, mit q^* aus

(26a) mit $\beta=1$, wenn $q^* < \infty$ ist. Kann man nur eine hinreichende Bedingung finden, so daß q^* mit $\beta=1$ aus (26a) tatsächlich endlich ist, so ist damit die Existenz des Grenzwertes $q(n), n \rightarrow \infty$ gezeigt. Nun existiert aber ein $q^* < \infty$ aus (26a) mit $\beta=1$ dann, wenn $E(Y) < 0$ ist, denn aus (26a) gilt:

$$0 = \int_{B_1} y dF(y, z) + \int_{B_2} (z - q^* + f^*) dF(y, z) - q^* P\{B_3\} \quad (31)$$

mit:

$$B_1 = \{y, z | y > \max(z - q^* + f^*, -q^*)\}$$

$$B_2 = \{y, z | z - q^* + f^* \geq y ; z - q^* + f^* > -q^*\}$$

$$B_3 = \{y, z | -q^* \geq \max(y, z - q^* + f^*)\}.$$

Die Mengen B_1, B_2, B_3 bilden eine Zerlegung. Aus (31) folgt dann

$$0 = \int_{B_1} y dF(y, z) + \int_{B_2} (z - f^*) dF(y, z) - q^* (P\{B_2\} + P\{B_3\})$$

Damit ist eine hinreichende Bedingung dafür, daß $q^* < \infty$ ist: $P\{B_2\} + P\{B_3\} > 0$.

Für diese Bedingung hinreichend ist nun wiederum, $E(Y) < 0$, denn wäre $P\{B_2\} = P\{B_3\} = 0$, so gilt aus (31):

$$0 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y dF(y, z) = E(Y) \quad \text{und das ist ein Widerspruch zu}$$

der Annahme $E(Y) < 0$. Also gilt aus

$$E(Y) < 0 \rightarrow P\{B_2\} + P\{B_3\} > 0 \rightarrow q^* < \infty.$$

Somit ist also $E(Y) < 0$ hinreichende Bedingung für die Existenz des Grenzwertes $q^* < \infty$. Existieren die Grenzwerte von $q(n), f(n), n \rightarrow \infty$, so ergeben sich q^* u. f^* nach dem Satz von Lebesgue, aus (25a,b) wie in (26a,b) angegeben. Der Zusatz über f^*, q^* garantiert die Eindeutigkeit der Lösung, da ansonsten z.B. für $P(Z \leq -f^*) = 0, -f^*$ auf eine Menge vom Maß null jeden bel. Wert $\leq \inf Z$ annehmen könnte. Der Zusatz ergibt: $-f^* = \inf Z$.

(b) Die optimale Politik ergibt sich für den Fall $n = N-i+1 < \infty$ aus dem Vergleich der Maximanden in (25a,b). So sind z.B. die Auszahlungen aus (25a):

$$0, \quad \text{wenn } a_1 = 0, \quad b_1 = 0$$

$$Y + \beta q(n-1), \quad \text{wenn } a_1 = 1$$

$$Z + \beta f(n-1), \quad \text{wenn } a_1 = 0, \quad b_1 = 1$$

Die optimale Politik besteht nun darin, jeweils das Maximum aus den drei Entscheidungsmöglichkeiten auszuwählen. So ist z.B.

$$a_1^* = 0, \quad b_1^* = 0, \quad \text{wenn } 0 \geq \max(Y + \beta q(n-1), Z + \beta f(n-1))$$

Für den Grenzwert $n \rightarrow \infty$ konvergieren $q(n-1), f(n-1)$ gegen f^*, q^* und somit ergeben sich dann die optimalen Wahl- bzw. Stoppregeln wie in (27) angegeben.

Zu beachten ist allerdings, daß für den Fall $\beta=1, E(Z), E(Y)=0$, obwohl die Grenzwerte f^*, q^* existieren, dennoch eine optimale Politik nicht immer existiert. So kann beispielsweise $-f^* = \inf Z$ sein. Wenn Z eine stetige Zufallsgröße ist, ist somit $P(b_k^*=0) = P(Z \leq -f^*) = 0$, also wird der Prozess praktisch nie abgebrochen. Diese Lösung für f^* ist nur möglich, wenn $E(Z) = 0$ ist. Damit wäre aber nach dieser Politik die Auszahlung ebenfalls null. Für den Fall $\beta=1, E(Z), E(Y) \leq 0$, Y, Z stetig kann man sich durch die Konstruktion einer ξ -optimalen Politik helfen, indem man z.B. setzt: $b_k^*=0$, wenn $Z \leq -f^* + \xi$, Für die anderen in Res.5(b) erwähnten Fälle ergeben sich keine Existenzprobleme.

Ein ähnliches Ergebnis kann man für Markovprozesse ableiten. Da wir im Zusammenhang mit Res.4 nur diskrete Zufallsgrößen zugelassen haben, ergeben sich keine Existenzprobleme für die optimale Politik. Indem wir in Res.4 eine Umindizierung vornehmen und für $v_1(y_{1-1})$ bzw. $m_k(z_{k-1})$ setzen $v_n(y')$ bzw. $m_n(z')$, $n=N-1+1, n=N-k+1$, und voraussetzen, daß der Prozess homogen ist (Typ F_{3a}), also alle Übergangswahrscheinlichkeiten von i unabhängig sind, können wir mit $\beta(i)=\beta$ das folgende Ergebnis sofort aus Res.4 ableiten:

Resultat 6:

(a) Für die Nettoprozesswerte im homogenen Markovprozess (F_{3a}) gilt bei $\beta < 1$

$\lim_{n \rightarrow \infty} v_n(y') := v(y') < \infty$ und $\lim_{n \rightarrow \infty} m_n(z') := m(z') < \infty$ existieren,

mit:

$$v(y') = \sum_{y \in I_1} \sum_{z \in I_2} \text{MAX}(y + \beta v(y), z + \beta m(z)) p(y|y') p(z), \quad y' \in I_1$$

$$m(z') = \sum_{z \in I_2} \text{MAX}(z + \beta m(z), 0) p(z|z'), \quad z' \in I_2,$$

mit dem Prozesswert zu Beginn ($i=1$):

$$v = \sum_{y \in I_1} \sum_{z \in I_2} \text{MAX}(y + \beta v(y), z + \beta m(z), 0) p(y) p(z),$$

(b) Die optimalen Politiken ergeben sich als:

$$a_1^* = 0, b_1^* = 0, \text{ wenn } 0 \geq \text{MAX}(Y + \beta v(Y), Z + \beta m(Z))$$

$$a_1^* = 1, \text{ wenn } Y + \beta v(Y) \geq Z + \beta m(Z) \text{ u. } Y + \beta v(Y) > 0$$

$$a_1^* = 0, b_1^* = 1, \text{ wenn } Z + \beta m(Z) > \text{MAX}(Y + \beta v(Y), 0)$$

$$b_k^* = 0 \iff Z \leq -\beta m(Z).$$

Das Ergebnis läßt sich ähnlich wie Res.5 auch auf den Fall ohne Diskontierung: $\beta=1$ ausdehnen, wenn man fordert, daß die (bedingten) Erwartungswerte von Y und Z nicht positiv sind.

4. Beispiel eines Produkt-Ersetzungs-Problems

Die Grenzbetrachtungen des vorausgegangenen Kapitels ermöglichen es nunmehr für einfache Beispiele eine Lösung explizit anzugeben. Wir gehen aus von dem Res.6, also einem Markov-Umschaltprozess und nehmen an, daß das Problem darin besteht ein altes Produkt (a) gegebenenfalls durch ein neues (b) zu ersetzen oder die Produktion beider Produkte (a), (b) bei schlechten Absatzchancen einzustellen.

Die Gewinne des Produktes (a) können nur jeweils 3 Zustände, nämlich y_1, y_2, y_3 annehmen, mit: $y_3 > y_2 > y_1$, die Gewinne des neuen Produktes (b) nur zwei Zustände z_1, z_2 ; $z_2 > z_1$. Die Anfangsverteilungen $P\{Y_1 = y_j\} = p_j$ und $P\{Z_1 = z_j\} = \tilde{p}_j$ seien gegeben, $j \in \{1, 2, 3\}$, bzw. $j \in \{1, 2\}$, und die Übergangswahrscheinlichkeiten $p(y_j | y_1) = p_{1,j}$, $j \in \{1, 2, 3\}$ bzw. $p(z_j | z_1) = \hat{p}_{1,j}$ seien durch die nachfolgenden Matrizen P bzw. \hat{P} bestimmt.

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1-p & p & 0 \end{bmatrix} \quad \hat{P} = \begin{bmatrix} s & 1-s \\ 1-r & r \end{bmatrix}$$

Für dieses Beispiel ergäbe sich aus Res.6 das folgende Lösungsproblem:

$$v(y_1) = \max(y_2 + \beta v(y_2), z_1 + \beta m(z_1), 0) \tilde{p}_1 \\ + \max(y_2 + \beta v(y_2), z_2 + \beta m(z_2), 0) \tilde{p}_2 \quad (32a)$$

$$v(y_2) = \max(y_3 + \beta v(y_3), z_1 + \beta m(z_1), 0) \tilde{p}_1 \\ + \max(y_3 + \beta v(y_3), z_2 + \beta m(z_2), 0) \tilde{p}_2 \quad (32b)$$

$$v(y_3) = (1-p) \max(y_1 + \beta v(y_1), z_1 + \beta m(z_1), 0) \tilde{p}_1 \\ + \max(y_1 + \beta v(y_1), z_2 + \beta m(z_2), 0) \tilde{p}_2 \\ + p \max(y_2 + \beta v(y_2), z_1 + \beta m(z_1), 0) \tilde{p}_1 \\ + \max(y_2 + \beta v(y_2), z_2 + \beta m(z_2), 0) \tilde{p}_2 \quad (32c)$$

$$m(z_1) = s \max(z_1 + \beta m(z_1), 0) + (1-s) \max(z_2 + \beta m(z_2), 0) \quad (32d)$$

$$m(z_2) = (1-r) \max(z_1 + \beta m(z_1), 0) + r \max(z_2 + \beta m(z_2), 0). \quad (32e)$$

Ohne zusätzliche Voraussetzungen wären für dieses Beispiel bei der Ausrechnung eine Reihe von Fallunterscheidungen zu treffen. Zur Demonstration soll im folgenden ein sinnvoller Fall herausgegriffen werden. Dazu wird vorausgesetzt, daß die optimalen Politiken bekannt sind und die Lösung für die Prozesswerte dann ausgerechnet. (Das numerische Beispiel kann dann entsprechend gewählt werden).

Die optimalen Produkt-Politiken seien gegeben durch:

$$a_1^* = 0, b_1^* = 0, \quad \text{wenn } y_1 = y_1 \quad \text{und} \quad z_1 = z_1 \quad (33a)$$

$$a_1^* = 1 \quad \text{wenn } y_1 = y_2 \quad \text{und} \quad z_1 = z_1 \quad (33b)$$

$$a_1^* = 0, b_1^* = 1, \quad \text{wenn } y_1 = y_2 \quad \text{und} \quad z_1 = z_2 \quad (33c)$$

$$b_k^* = 0 \quad \longleftrightarrow \quad y_k = y_1. \quad (33d)$$

Also wird der Gesamtprozess - sowohl Produkt (a) als auch (b) - gestoppt vor der Periode 1, wenn jeweils die niedrigsten Gewinne (Verluste) y_1 und z_1 eintreten. Das Produkt (a) wird weiter produziert in der Periode 1, wenn der Gewinn von (a) relativ hoch ist (y_3 oder y_2) und der Gewinn von (b) niedrig (z_1). Es wird in der Periode 1 das Produkt (a) ersetzt durch (b), wenn der Gewinn von (a) relativ niedrig (y_2 oder y_1) ist und der von (b) relativ hoch (z_2).

Für diese Politik ergibt sich aus (32):

$$m(z_2) = \frac{rz_2}{1-\beta}, \quad m(z_1) = \frac{z_2(1-s)(1+\beta r-\beta)}{1-\beta}, \quad v(y_2) = y_3 + v(y_3),$$

$$v(y_1) = \tilde{p}_1\{y_2 + \beta v(y_2)\} + \tilde{p}_2\{z_2 + m(z_2)\}$$

$$v(y_3) = \frac{p\tilde{p}_1(y_2 + \beta y_3) + \tilde{p}_2(z_2 + \beta v(z_2))}{1 - p\beta^2}$$

Für das numerische Beispiel: $y_1 = -30, y_2 = -5, y_3 = 15, z_1 = -25, z_2 = 10$, $p = 0,5, r = 0,2, s = 0,4, \tilde{p}_1 = 0,4, \tilde{p}_2 = 0,6, \beta = 0,9$ ergibt sich:
 $m(10) = 20, m(-25) = 22,8, v(-5) = 41,9, v(-15) = 29,8, v(15) = 30,9$.

In der Praxis müssen die Rechnungen im Rahmen eines Computermodells vorgenommen werden, so daß auch der Einfluß von Datenänderungen mittels Simulation leicht festgestellt und zur Entscheidungshilfe herangezogen werden kann.

Literatur:

- [1] Henke, M. Produkt-Abbruch-Entscheidungen bei Unsicherheit
in: Zeits. für Betriebswirtschaft, 1972
- [2] Henke, M. Zum Stoppen von Forschungs- und Entwicklungs-
projekten, in: Zeits. für ges. Staatsw., 128,1,1972,
S.39-64

Prospective Planning. A Practical Application in Defense

by W. K. Brauers, Brüssel

1. ALTERNATIVE THINKING

From economics we know that the number of human wants is extremely large but that the means are limited. If we project this statement into the future, we may say that future wants are limited by the means, not only by the future means, but also by the actually available means. Even stronger : actual means help to determine the future means.

The actual national wealth and the increase in the actual national wealth, determine the national wealth of the future. The rise of the Chinese Democratic Republic in the future is limited by its actual national wealth and by its economic growth rate.

The term "automation" is more than thirty years old (1), but automation is still very limited in certain fields. In the United States out of 2, 175, 309 machines only 13,689 are with all tape control (2). Productivity do not increase with more than 3 % a year in the United States. These statements explain the wrong point of view of some that the distant future is easily determined by to-day events.

Other people will take the opposite point of view : the future and certainly the far distant future is not to foresee.

The famous economist, Lord J.M.KEYNES, said once : "on the long run we are all dead". Who could foresee the rise of the Soviet economy fifty years ago ?

(1) L.LANDON GOODMAN, Man and Automation, Harmondsworth, 1957.

(2) The American Machinist, Tenth Inventory, 1968.

In the industrial age there was always a long distance between invention at one side and industrial application and economic multiplication at the other side. There was a long distance between the invention of photography and its economic multiplication. Even plastics have known their invention period in the thirties against their real success only after the second world war. Uranium was of no value in 1940 and five years later it was applied in the atomic bomb, but industrial application wanted another ten years. Recently, there was however a speeding up of technological progress. This was the case for a large part of the electronics industry (1). The industrial structure of the Western European countries changes thoroughly in a period of approximately five years. All this explains the scepticism of some in connection with futures research.

What are our comments at one side on some who find that the distant future is exclusively determined by to-day events, at the other side on the others who are sceptical about future events ?

(1) First of all : the factor war is a factor of uncertainty for the distant future. Peace research may help to know this factor better.

(2) Secondly : if the range of application of futures research is smaller than this of the world then the diversion of means is easier and ipso facto the outcome more uncertain. E.g. futures research about the electronics industry in Belgium in the year 2.000 is much more difficult than for the same industry in the world. War preparation is a more uncertain factor for Sweden than for the U.S.

(1) Vide : J.J. SERVAN-SCHREIBER, Le défi américain, Paris 1967, p. 76.

(3) Thirdly : in my opinion the uncertainty about the distant future is decreased through the application of alternative thinking : one tries to find several alternative possibilities for a future event. One applies something like a decision-tree but this time for possible outcomes on the long run. Indeed, instead of the extrapolation method (fig. 1), in which goal a is aimed at while in reality the not foreseen goals b, c and d are possible, one prefers the following scheme (fig. 2) (1) :

(1) The figures may give the false impression that it only concerns about linear extrapolations. In fig. 1 as well as linear extrapolation as for example an exponential or a logistic curve may be involved. In fig. 2 also a graphic relationship is possible between the different alternative goals. LIVINGSTON cited the case of the enveloping curve of the number of units accelerator energy over time which encloses the different alternative productions in the generators, the cyclotron, the betatron and the synchrotrons (M.S. LIVINGSTON, Introduction to the Development of High Energy Accelerators, Dover Publications Inc., New York 1966). The difficulty lies in the fact that located on a certain curve one is not sure about a possible enclosing by an enveloping curve or about its development.

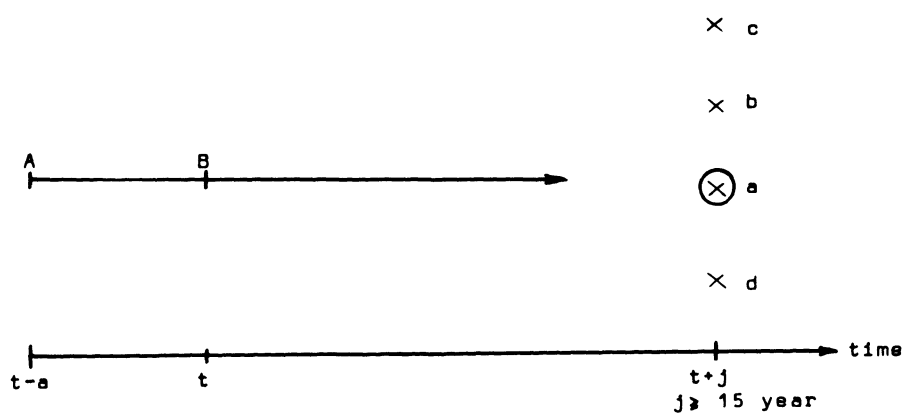


Fig. 1
The extrapolation - method

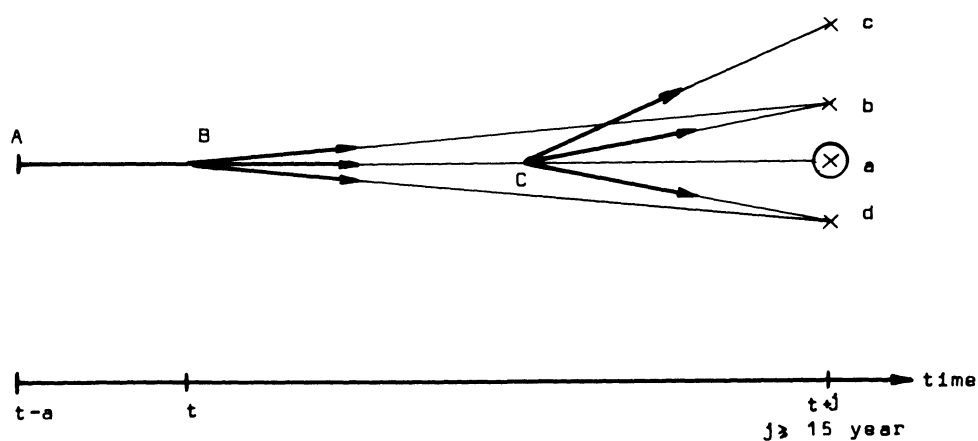


Fig. 2
Decision - tree under uncertain circumstances

In B of figure 2 goals a, b and d are visualized as possible. When the period considered is expired b, c and d are possible, but c is not foreseen and a is unrealistic.

In point C however, better insight already exists into the future than at B, and this time b, c and d are visualized as possible goals. This means a pledge for flexibility and revision of the system.

Alternative thinking is a kind of philosophy which we have to acquire. Out of the milieu-studies of national defense we give two examples of alternative thinking :

- (1) Firstly, against the possible world aggressive policy of China in the XXIth century two ways are open, either to bring world wants and means into accordance with one another and to distribute them geographically, or to erect a new "Chinese wall" around Europe in order to protect it in a military way.
- (2) Secondly, to see Europe's military future, either in the formation of a European Army, or the revision of NATO, or the neutralization of Europe.

2. PROSPECTIVE THINKING AND PROSPECTIVE PLANNING

Prospective thinking distinguishes itself from extrapolation, prognosis and prediction by its imagination about the long run (e.g. 15 years or more) ; it distinguishes itself from futurology by the introduction of alternative thinking.

We are however more interested in the application than in more passive thinking. Prospective thinking may not degenerate in Hamletian doubts. What can actively generate from prospective thinking ?

It is activated by prospective planning which depicts the possible results of alternative solutions for the long run (1). On a certain moment, viz. in point B of figure 2, we have even to decide what direction we have to go, if our present attitude is affected with it. It can still be possible to postpone any choice in B, but it is also possible that the economic costs of this attitude becomes prohibitive if one has e.g. to make expenditures which have to bring us either to b or to d. In figure 3 above the decision is postponed till B' ; this however increases economic costs to a considerable extent, as depicted in figure 3 underneath. Costs would have been much lower if one had immediately made a choice in point B between solutions b and d.

All this creates a new alternative solution in B : to postpone any choice.

The discovery and description of alternative solutions towards the fulfillment of a certain goal is eased through techniques such as "delphi", "cross impact analysis", "simulation" and "gaming".

Finally, the decision maker is assisted in his choice by "cost-effectiveness models", "sensitivity analysis", "cross impact analysis", "gaming", "simulation" and similar techniques.

(1) We consider planning as a neutral, scientific technique, which is distinct from the structure of the nation in which it is used. This structure plays however a role in the enforcement of the planning in which sense one speaks of indicative, orientated and imperative planning.

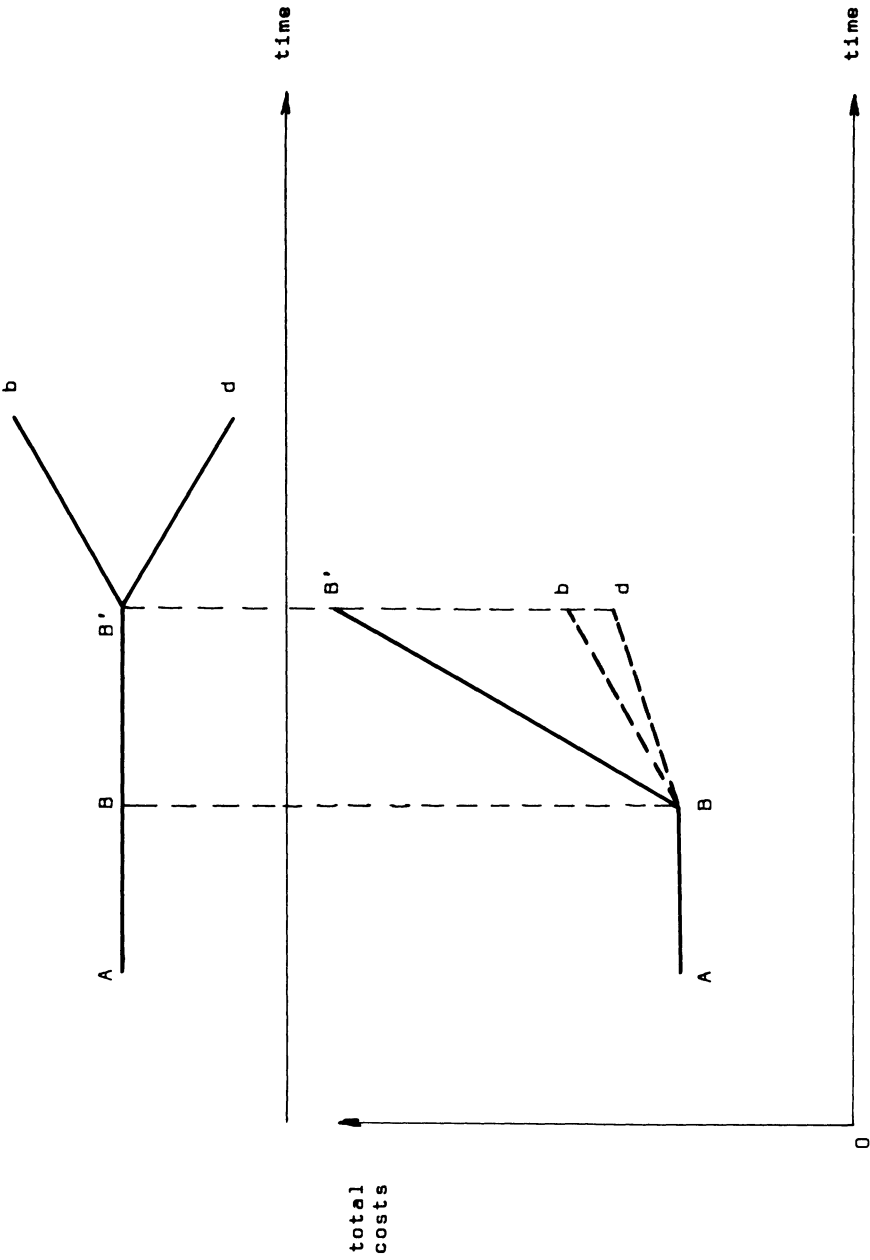


Fig. 3
The postponement of the decision

HOW IS PROSPECTIVE PLANNING LINKED TO OTHER FORMS OF PLANNING ?

Planning for the half-long term is traditionally produced in an enterprise or nationwide. This planning tries to bring a realistic picture of the half-long term future and this future is consciously or unconsciously for the largest part determined by to-day events and possibilities.

We saw that the starting point of prospective planning is different : in the framework of the imagination about the future of prospective thinking, prospective planning depicts the possible results of alternative solutions for the long run. As the starting point is different, antagonism is possible between prospective planning and planning for the half-long term. Such antagonism is already known between the existing planning for the half-long term and the budgeting (1).

Budgeting in more or less detail depicts the management decisions to be taken on the short term of one year. Budgeting is strongly linked to the situation of to-day, especially to the day-to-day expenditures.

In a system of government management originated in the United States and called P.P.B.S. (Planning, Programming, Budgeting System) one tries to bring the planning for the half-long term in accordance with the budgeting in an operation which is called the programming process. In our system the bridge firstly between prospective planning and planning for the half-long term and secondly between planning for the half-long term and budgeting is each time made by the programming process.

The programming between the planning for the half-long term and the budgeting mainly consists of an option between capital and current expenditures.

(1) In the economic planning a similar antagonism is known between growth models at one side and cyclical models at the other side ; this created the need for the development of so-called 'cyclical-growth models'.

With capital expenditures one means these for investments and for consumption durables, with current expenditures these for personnel, operation and maintenance.

The programming between the planning for the half-long term and the prospective planning has to find a harmonious solution between the expenditures for both kind of planning. Planning for the half-long term puts the accent on the urgent needs on the half-long term i.e. on the urgent capital expenditures, mainly replacement expenditures. Prospective planning will rather ask expenditures for scientific research and development and for the investments which are necessary for the realization of new projects. There is still another jump between the prospective planning and the planning on the half-long term at one side, the budgeting at the other side, viz. that the budget is calculated in real money costs, the plans however also in social costs and in cost-effectiveness terms (1). The experts of the programming as well as these of the prospective planning have to demonstrate the existing alternative solutions and to make based on scientific reasoning propositions for an option. The final decision however remains with the policy-maker (politician, manager).

In this way, the following scheme is interesting :

(1) The calculation in social costs may be less necessary for enterprises. In the future however, enterprises have to learn more and more what social costs they lay to the charge of the community by a certain production.

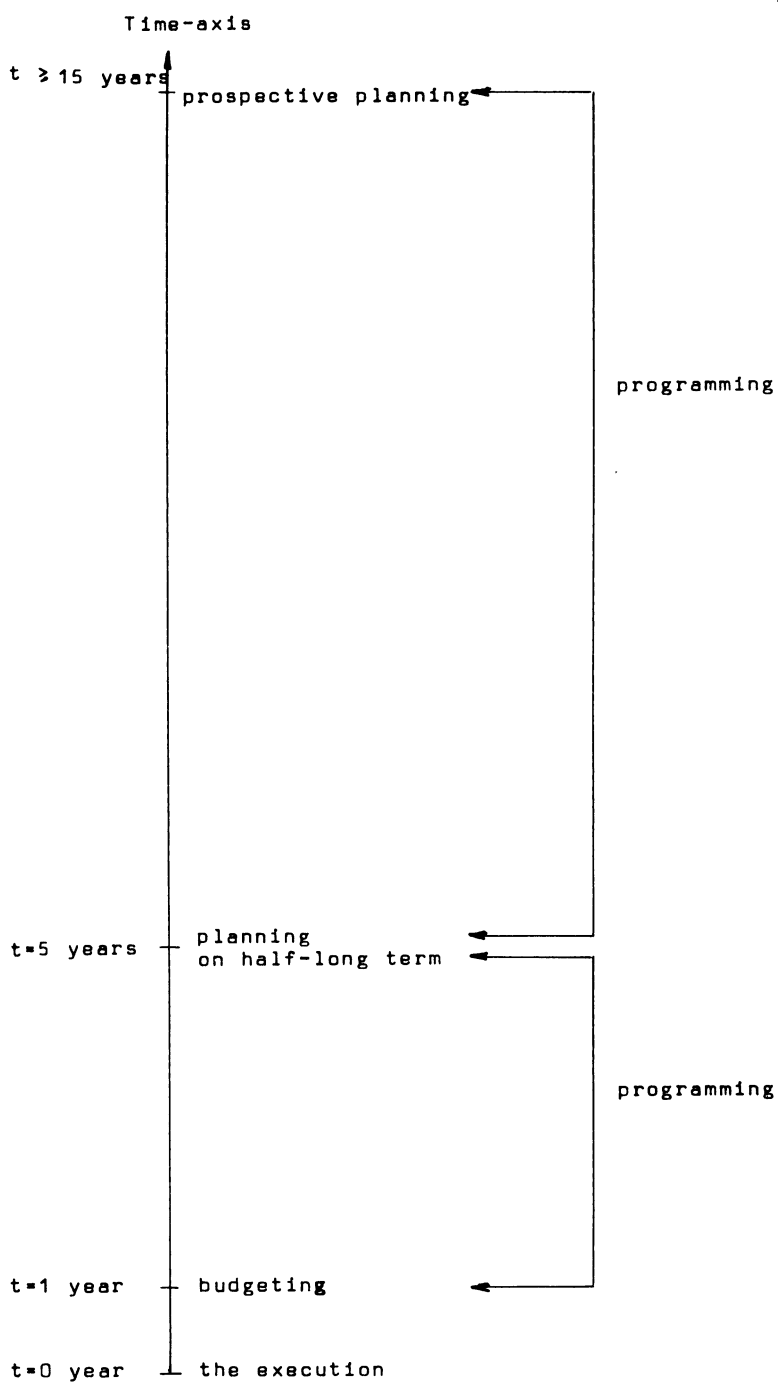


Fig. 4

The enlarged P.P.B.S.

Consequently, the "Planning, Programming, Budgeting System" becomes an enlarged system including prospective planning. This new system is not only applicable for government services but also for all private enterprises and even for all institutions, communities and individuals.

For long time it was said that government services are missing any rational economic functioning as the goal of profit making is lacking. Recently was maintained that even in private enterprises the ultimate goal of profit-making is obscured by sub-goals or side-goals. In any case the goal of profit making is not essential for private enterprises in the long run. It means that in the long run, as well as for private enterprises as for government services, a search for goals is necessary and that also for private enterprises prospective planning with alternative solutions is important.

3. PROSPECTIVE PLANNING IN DEFENSE

The probability of an attack by a potential enemy, the search for and the diminishing of the causes of warfare are the tasks of the "Peace-Research". For national defense however, the enemy attack has to be taken for granted. So in defense we have not to calculate the risk for war, but we have to prepare for war. For defense the attack is a donnée, of which the characteristics and size have to be studied.

Prospective planning in defense is then possible with as ultimate goal : to ensure national security by repelling any attack. For the nation as a whole this problem of defense only concerns partial planning, as well as e.g. the planning of the construction of safe roads in the prospective planning of transportation. Partial prospective planning like this for defense or transportation has to be subordinated to the "overall" prospective planning of the nation (1).

(1) For other examples of partial planning, see : H.S.BECKER and R. DE BRIGARD, A Framework for Community Development Action Planning, Institute for the Future, Middletown (Conn.) N° R - 18 and 19, 1971.

Once the goal known, how does prospective planning work in defense ?

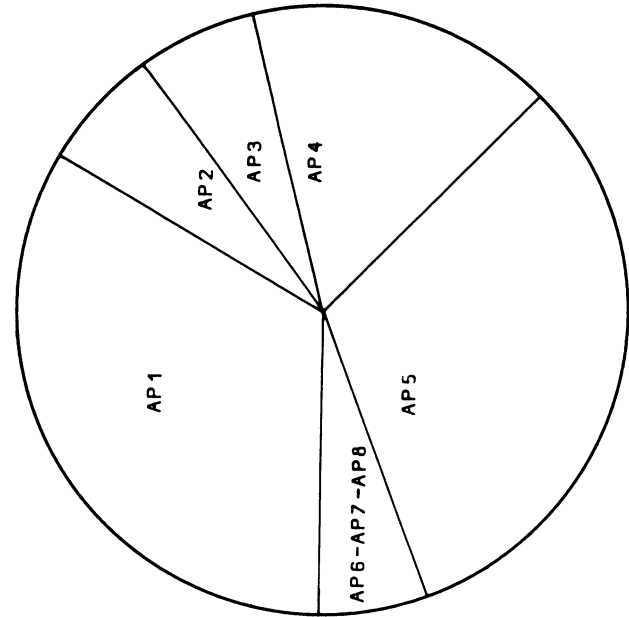
First of all the starting point is the environment in which the security and defense policy of the country should be defined. It mainly concerns the international environment. Therefore milieu-studies are made : "scenario writing" of alternative possibilities in the evolution of the international relations. In fact these "milieu-studies" should be the collective work of politicians, functionaries of the department of defense, higher officers of all the forces, functionaries of the department of foreign affairs and of independent scientific researchers and professors.

Assistance therefore is given by such techniques as delphi, cross impact, simulation and gaming. Then prospective planning in defense can start : a good case-study of what prospective thinking has to be in general.

Prospective planning in defense is illustrated here with the example of Sweden, a country which considers the matter as non-confidential and which possesses a policy of complete non-alignment in peace aiming at neutrality in war i.e. which has to think and act about defense all by itself. Sweden designed a prospective planning for 15 years (1). This period is chosen as the important military equipment generally possesses a lifetime of less than 15 years. Each four years the prospective planning is completely revised. The planning on the half-long term is formed by five year plans and then there is the budgetary year. Sweden, a small nation tries to put the size of the defense effort a bit higher than the opportunity cost for a possible aggressor, which is considered as a supreme power in a bi-supreme power-world. Till around 1987 nine cases of aggression are accepted as politically relevant for defense planning. Against these nine cases of aggression six defense responses are possible belonging to two main categories of alternative solutions (structures A and B) :

(1) ÖB 71, Perspektivplan 1972-1987, Preliminär öppen version, Försvarsstabens press-och upplysnings-avdelning, Stockholm 1971.
 ÖB 71, Bildbilaga, idem
The Defence Planning System, 4 vol. Ministry of Defence
 Sweden, Fö N° 1-4, 1970.

Structure B



Structure A

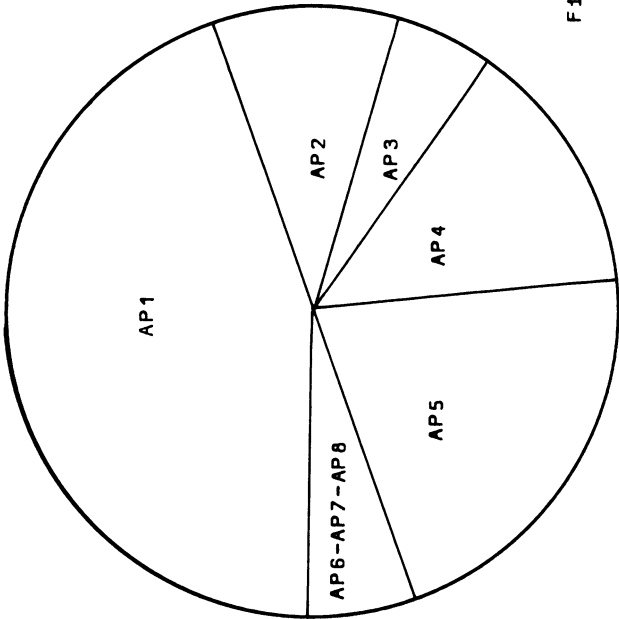


Fig. 5

Repartition of the means between the missions

- a. Structure A : the integral defense of the territory yard by yard ;
- b. Structure B : the forward defense.

In each of these categories of alternative solutions three alternative responses exist. The responses are elaborated by missions and not by services. In total eight missions are distinguished. The size of three missions is however not influenced by the choice of the response, it concerns :

AP 6 : central and regional staffs ;

AP 7 : common defense research ;

AP 8 : central agencies and functions.

Otherwise the available means are distributed between the missions after the chosen main category of alternative solutions in the following way (see fig. 5) :

As said, three responses belong to each main category of alternative solutions. These responses are a function of the money available i.e. each third response asks less money than the second response, the second response less than the first.

Per mission the composition of the weapon systems is as follows :

TABLE I - COMPOSITION OF THE WEAPON SYSTEMS PER MISSION IN THE PROSPECTIVE PLANNING OF SWEDISH DEFENSE

1. The Tactical Force (AP1) which is relatively much more important in structure A than in B (fig. 5).

Structure A is the traditional structure

Weapon systems	Response 1	Response 2	Response 3
1. Artillery regiments	-	-	15
2. Infantry brigade	43	51	50
3. "Norrlandsbrigader" (a)	21	16	13
4. Tankbrigades	36	33	22
	100	100	100

(a) equipped with tracked vehicles for the North.

Structure B is a complete new structure with modern units equipped with the most up-to-date weapons.

Weapon systems	Response 1	Response 2	Response 3
1. Mechanized brigades	87	90	88
2. Tankbattalions	13	10	12
	100	100	100

2. The Local Forces (AP_2) which are relatively much more important in structure A than in structure B (fig. 5).

Weapon systems	Structure A			Structure B		
	1	2	3	1	2	3
Local forces (a)	52	55	59	70	62	48
Fixed coast artillery	48	45	41	30	38	52
	100	100	100	100	100	100

(a) Enlisted on the spot from the age of 35 till 47

3. The Maritime Forces (AP_3) which this time are relatively more important in structure B than in structure A (fig. 5).

Structure A

Weapon systems	Resp.1	Resp.2	Resp.3
Speedboats with teleguided rockets	56	67	86
Gunboats	27	17	-
Helicopters	17	16	14
	100	100	100

Structure B

Weapon systems	Resp.1	Resp.2	Resp.3
Corvettes	6	-	-
Flotilla (a)	7	14	6
Speedboats with teleguided rockets	50	56	67
Gunboats	18	9	12
Helicopters	19	21	21
	100	100	100

(a) A small fleet

4. The Strategic Forces (AP_4) which are relatively much more important in structure B than in structure A (fig. 5).

Weapon systems	Structure A			Structure B		
	1	2	3	1	2	3
Light Assault planes	25	23	30	12	13	15
Heavy Assault planes	56	57	50	59	56	55
Submarines	19	20	20	29	31	30
	100	100	100	100	100	100

5. The Air Defense (AP_5) which is relatively much more important in structure B than in structure A (fig. 5).

Weapon systems	Structure A			Structure B		
	1	2	3	1	2	3
Interceptors	78	80	71	85	83	80
Missiles (fixed or mobile and of any size)	22	20	29	15	17	20
	100	100	100	100	100	100

4. SUMMARY

It concerns an original system of long-range planning to be used as well as for public economics as for corporate economics.

Prospective thinking distinguishes itself from extrapolation by its imagination about the future ; it distinguishes itself from futurology by the introduction of alternative thinking.

Prospective thinking is activated by prospective planning which depicts the possible results of alternative solutions. This prospective planning is assisted by such techniques as decision trees about the future outcome, delphi techniques and cross impact analysis, simulation and gaming. Prospective planning is a neutral and scientific method . The final decision remains in the hands of the policymaker (politician, manager). Prospective planning is linked up with the actual available means through an enlarged "Planning-Programming-Budgeting System".

The whole system is illustrated by a practical application in the defense of Sweden.

The Swedish example gives us the opportunity for a standardized proposition on terminology what prospective planning is concerned.

There are goals and subgoals or objectives ; the second is not only a part of the first ; it is more than that. There exists a hierarchical goal structure in the form of a pyramid going down with increasing specificity and decreasing permanency. "Since goals cover long time spans, it is useful to divide them into more specific subgoals that are more tangible to a community and require less time to accomplish. These subgoals, which should be viewed as means for achieving the larger goal, are usually called objectives" (1).

(1) H.S.BECKER and R. de BRIGARD, Op. cit., vol I, p.8

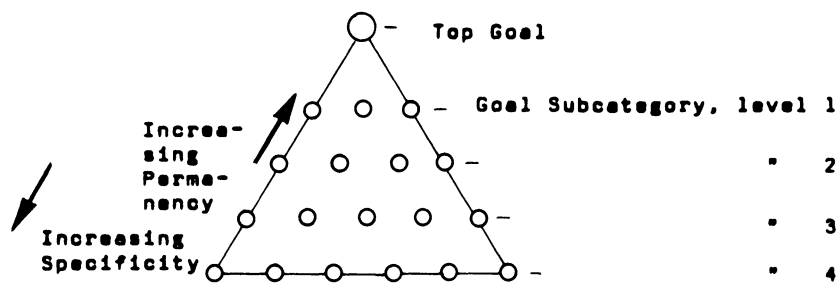


Figure 8 - A HIERARCHICAL GOAL STRUCTURE

What national defense is concerned the goal could be : to ensure the security of a country X and to contribute to the maintenance of World Peace. The subgoals or objectives however would e.g. rather be :

- 1° the military defense of a country in case of attack ;
- 2° protection of the country against threats other than those of foreign military attack by surveillance and control of the national land, sea and airspace ;
- 3° contribution to a collective defense under a Treaty such as NATO ;
- 4° contribution to international stability in the world ;
- 5° national support etc.

The goal subcategory, level 2 as shown in fig. 6 is called in defense analysis : a mission.

So e.g. in the military defense of Sweden eight missions were defined.

In the contribution to a collective defense under the NATO-Treaty, Belgium has to fulfill the following missions :

- 1° the Belgian Army brigades in the German Federal Republic under the Northern European Command have to defend a sector
- 2° the Belgian Air Force is a part of the Second Allied Tactical Airforce which has to collaborate with the Northern Army Group ;
- 3° the Belgian Navy participating in the NATO Channel Committee has to take care of a sector of the North Sea and of the Belgian territorial waters.

In the contribution to international stability the following missions are e.g. distinguished : peacekeeping operations, truce supervisory operations and military training assistance. In the national support the following missions are e.g. possible : emergency and disaster, search and rescue.

The goal subcategory level 3 in defense is formed by the targets of the weapon systems. E.g. fixed coast artillery is a weapon system of the mission "local forces".

By activity is the activity meant of a weapon system, e.g. the activity of the missiles in the mission of the air defense trying to hit their targets. The target is the objective of a weapon system.

Weapon systems and missions are the building blocks for a strategy. There exists a choice of strategies either in attack or in response. We saw that in the Swedish case there are two broad categories of responses on a possible attack : the rear defense and the forward defense. In each broad category several variants exist after the size of the intended expenditures.

We saw earlier that in the non-defense field one speaks of alternative possibilities, alternative solutions, rather than of alternative strategies. It is the essence of prospective thinking and planning.

Lagerhaltung

**Die stationäre Behandlung von Mehr-Produkt
Lagerhaltungsmodellen**
von D. Hochstädter, München

Zusammenfassung:

Für ein Zwei-Produkt/Ein-Lager Modell werden verschiedene Formen einer Bestellpolitik definiert. Für den Fall, daß zukünftige Kosten nicht diskontiert werden, wird auf der Basis der erwarteten, durchschnittlichen Kosten pro Periode ein Ausdruck für diese Kosten erhalten. Aus diesem Ausdruck können die optimalen Parameterwerte der vorgegebenen Politik mit den Methoden der Differentialrechnung erhalten werden. Die Ergebnisse lassen sich ohne Schwierigkeiten formal auf Mehr-Produkt Modelle übertragen.

- 1) Referat, gehalten auf der 10. Jahrestagung der DGU, Bochum, 22.-24. September 1971
- 2) Privatdozent Dr. Dieter Hochstädter, Institut für Angewandte Mathematik der TU München, 8 München 2, Arcisstr. 21

1. Einleitung

Das Interesse an der Theorie der Lagerhaltung hat sich in den letzten Jahren vom Ein-Produkt/Ein-Lager Modell den Mehr-Produkt/Ein-Lager bzw. Ein-Produkt/Mehr-Lager Modellen zugewandt. Durch Übertragung der beim Ein-Produkt/Ein-Lager Modell verwandten Techniken war man im begrenzten Rahmen in der Lage, Aussagen über die qualitative Form der optimalen Politiken für verschiedene Modellsituationen zu geben. Diese haben gewisse Ähnlichkeiten mit denen für das Ein-Produkt/Ein-Lager Modell bekannten optimalen Politiken. Die bisher bekannten Untersuchungen, über die der Autor in [3] berichtet hat, bieten jedoch wenig Anhaltspunkte zur Entwicklung von Rechenverfahren, die es ermöglichen würden, die Parameter einer optimalen Politik mit einem für die praktische Anwendung vertretbaren Rechenaufwand zu bestimmen. Ein Algorithmus, der sich bei der Bestimmung der Parameter einer optimalen (s, S) -Politik für das Ein-Produkt/Ein-Lager Modell als recht effizient erwiesen hat [5], ist von Veinott und Wagner [7] entwickelt worden. Dieser benutzt Ideen, die aus der Erneuerungstheorie stammen.

In der vorliegenden Arbeit sollen einige Gedankengänge dieses Algorithmus auf ein Zwei-Produkt/Ein-Lager Modell übertragen werden. Für verschiedene Formen der Bestellpolitik kann man auf der Basis der durchschnittlichen Kosten pro Periode die zugehörige Kostenfunktion aufstellen und zeigen, daß ihre Lösung auf einfache Beziehungen führt, die formal mit den Ergebnissen des Ein-Produkt/Ein-Lager Modells übereinstimmen. Der Fall, daß zukünftige Kosten auf ihren heutigen Wert diskontiert werden, wurde vom Autor an anderer Stelle behandelt [4]. Die numerische Bestimmung der optimalen Parameter läßt sich dann rekursiv durchführen. Eine formale Erweiterung dieser Resultate auf mehr als zwei Produkte ist durch Einführung von Vektoren möglich.

2. Das Ein-Produkt/Ein-Lager Modell

Es werden folgende Annahmen getroffen: Entscheidungen können nur zu Beginn vorgegebener Inspektionszeiträume (Perioden) getroffen werden. Die Lagerauffüllung erfolge ohne Lieferverzögerung. Die Nachfrage nach dem Gut werde in jeder Periode durch eine nicht-negative Zufallsvariable U bestimmt, deren Dichtefunktion $f(u)$ bekannt sei. Die Nachfrage aufeinanderfolgender Perioden werde durch eine Folge unabhängiger und identisch verteilter Zufallsvariabler bestimmt. Die während einer Periode auftretenden Fehlmengen können zu einem späteren Zeitpunkt nachgeliefert werden. Es werden die folgenden, zum Teil üblichen, Bezeichnungen verwendet:

x := Lagerbestand vor der Entscheidung,
 y := Lagerbestand nach der Entscheidung,
 z := $y - x$:= Bestellmenge,
 $c(.)$:= proportionale Bestellkosten,
 $h(.)$:= Lagerungskosten,
 $p(.)$:= Fehlmengenkosten,
 K := fixe Bestellkosten,
 $L(y)$:= Ein-Perioden Kosten bei einem Anfangsbestand von y Einheiten, es gilt

$$L(y) = \begin{cases} (h+p) \int_{-\infty}^y (y-u)f(u)du - p(\mu-y) & \text{für } y \geq 0, \\ p(\mu-y) & \text{für } y < 0, \end{cases}$$

wobei $\mu = \int_{-\infty}^{\infty} uf(u)du$ den Erwartungswert der Nachfrageverteilung bezeichnet.

(s,S) -Politik := Für jeden Inspektionszeitpunkt gilt
 falls $x < s$ bestelle $z = S-x$,
 falls $x \geq s$ bestelle $z = 0$.

$C^n(x)$:= Erwartungswert der gesamten, entscheidungsabhängigen Kosten eines n -Perioden Problems bei einem Anfangsbestand von x Einheiten, wenn eine (s,S) -Politik befolgt wird.

Es werde nun angenommen, daß zukünftige Kosten nicht auf ihren heutigen Wert diskontiert werden. Es soll zunächst ein Ausdruck für $C^n(x)$ gefunden werden. Mit $x(t)$ und $y(t)$ werden der Lagerbestand vor bzw. nach der Entscheidung zu Beginn der Periode t bezeichnet. Es gilt

$$y(t) = \begin{cases} S & \text{falls } x(t) < s, \\ x(t) & \text{falls } x(t) \geq s, \end{cases}$$

sowie

$$x(t+1) = y(t) - u(t).$$

Die Kosten $C^n(x)$ werden gegeben durch

$$C^n(x) = \sum_{t=1}^n E\{K \delta[y(t)-x(t)] + L[y(t)] \mid x(1) = x\},$$

mit $\delta(0) = 0$ und $\delta(z) = 1$ für $z > 0$.

Mit T_x werde diejenige Periode bezeichnet, in der die kumulierte Nachfrage erstmals die nicht-negative Zahl x übersteigt, d.h.

$$T_x = \min_n (n \mid S_n \geq x, x \geq 0).$$

Ferner werde für $x \geq s$ die Größe $T_{x-s}(k)$ definiert als

$$\begin{aligned} T_{x-s}(k) &= \text{Prob}(T_{x-s}=k) = \text{Prob}(T_{x-s} \leq k) - \text{Prob}(T_{x-s} \leq k-1) \\ &= \text{Prob}(S_k > x-s) - \text{Prob}(S_{k-1} > x-s) \\ &= F^{(k-1)}(x-s) - F^{(k)}(x-s) \end{aligned}$$

für $k \geq 1$ und $T_{x-s}(0) = 0$.

Somit wird durch die Größe $T_{x-s}(k)$ die Wahrscheinlichkeit gegeben, daß die kumulierte Nachfrage erstmals in der Periode k den Betrag $x-s$ übersteigt.

Es folgt unmittelbar, daß für $x \geq s$ ebenfalls gilt

$$\sum_{k=0}^n T_{x-s}(k) = 1 - F^{(n)}(x-s) \quad \text{für } n \geq 1, \text{ und}$$

$$\sum_{k=0}^n k T_{x-s}(k) = 1 + \sum_{k=1}^{n-1} F^{(k)}(x-s) - n F^{(n)}(x-s) \quad \text{für } n \geq 1.$$

Es gilt

$$\lim_{n \rightarrow \infty} F^{(n)}(x) = 0, \text{ und}$$

$$\lim_{n \rightarrow \infty} \sum_{u=1}^n F^{(u)}(x) = M(x),$$

für jedes $x \geq 0$, wobei

$$M(x) = \sum_{n=1}^{\infty} F^{(n)}(x),$$

für $x \geq 0$, die Erneuerungsfunktion bezeichnet. $M(x)$ kann interpretiert werden als diejenige Zahl von Perioden, die verstreicht, bevor die kumulierte Nachfrage den Wert x übersteigt.

Es folgt unmittelbar

$$\sum_{k=0}^{\infty} T_{x-s}(k) = 1, \text{ und}$$

$$\sum_{k=0}^{\infty} k T_{x-s}(k) = 1 + M(x-s).$$

Setzt man $T_D^{(1)}(k) = T_D(k)$, mit $D = x-s$, dann gilt

$$T_D^{(n)}(k) = \sum_{u=0}^k T_D^{(n-1)}(u) T_D(k-u)$$

für $k \geq 0$ und $n \geq 2$. Man kann die Größe $T_D^{(n)}(k)$ als die Wahrscheinlichkeit interpretieren, daß die kumulierte Nachfrage den Betrag D in der Periode k zum n -ten Male übersteigt.

Ferner werde die Größe

$$t_D^{(n)}(k) = \sum_{n=1}^{\infty} T_D^{(n)}(k)$$

für $k \geq 0$, mit $t_D(0) = 0$, definiert.

Wenn $x(1) = x < s$ ist, dann wird $y(1) = S$ und es gilt

$$C^n(x) = K + C^n(S), \quad \text{für } x > s \text{ und } n \geq 1.$$

Für $x(1) = x \geq s$ wird die Wahrscheinlichkeit, daß $t = k+1$ die erste Periode ist, in der der Bestand $x(t)$ unter den Schwellenwert s absinkt, d.h. $x(t) < s$ ist, durch $T_{x-s}^{(k)}$ gegeben.

Man erhält [2]

$$C^n(x) = L(x) + \sum_{k=1}^{n-1} \sum_{u=0}^{x-s} L(x-u) f^{(k)}(u) + \sum_{k=1}^{n-1} \{K + C^{n-k}(S)\} T_{x-s}^{(k)}$$

für $n \geq 1$ und $x \geq s$, wobei $f^{(k)}(u)$ die k -fache Faltung der Nachfragedichte $f(u)$ bezeichnet.

Da jedoch der Gesamterwartungswert dieser Kosten für n gegen unendlich nicht gegen einen Grenzwert strebt, kann man auch nicht von einem Gesamterwartungswert der Kosten des Problems sprechen. Um jedoch auch für das unendliche Modell endliche Kosten zu finden, die man optimieren kann, zerlegt man die Kosten in Durchschnittskosten g pro Periode und in ein Restglied $g^n(x)$,

$$C^n(x) = ng + g^n(x) \quad \text{für } n \geq 0.$$

Sodann bestimmt man einen Ausdruck für $g^n(x)$ und erhält nach einigen Umformungen

$$\begin{aligned} g^n(x) = & L(x) + \sum_{k=1}^{n-1} \sum_{u=0}^{x-s} L(x-u) f^{(k)}(u) + K \sum_{k=1}^{n-1} T_{x-s}^{(k)} \\ & + \sum_{k=1}^{n-1} \{g^{n-k}(S) + (n-k)g\} T_{x-s}^{(k)} - ng \end{aligned}$$

$$\begin{aligned}
&= L(x) + \sum_{k=1}^{n-1} \sum_{u=0}^{x-s} L(x-u) f^{(k)}(u) + K \left[1 - F^{(n-1)}(x-s) \right] \\
&\quad - g \left[1 + \sum_{k=1}^{n-1} F^{(k)}(x-s) \right] + \sum_{k=1}^{n-1} g^{n-k}(s) T_{x-s}(k).
\end{aligned}$$

Setzt man zur Abkürzung

$$\begin{aligned}
b^n(x) &= L(x) + \sum_{k=1}^{n-1} \sum_{u=0}^{x-s} L(x-u) f^{(k)}(u) + K \left[1 - F^{(n-1)}(x-s) \right] \\
&\quad - g \left[1 + \sum_{k=1}^{n-1} F^{(k)}(x-s) \right],
\end{aligned}$$

$$b^0(x) \equiv 0,$$

für $x \geq s$, so erhält man die Erneuerungsgleichung

$$g^n(x) = b^n(x) + \sum_{k=0}^n g^{n-k}(s) T_{x-s}(k) \quad \text{für } x \geq s \text{ und } n \geq 0.$$

Speziell für $x = s$ erhält man

$$g^n(s) = b^n(s) + \sum_{k=0}^n g^{n-k}(s) T_D(k), \quad n \geq 0$$

mit der Lösung [1]

$$g^n(s) = b^n(s) + \sum_{k=0}^n b^{n-k}(s) t_D(k),$$

Durch Kombination erhält man für die Kostenfunktion

a. $x < s$:

$$C^n(x) = K + ng + b^n(s) + \sum_{k=0}^n b^{n-k}(s) t_D(k),$$

b. $x \geq s$:

$$C^n(x) = b^n(x) + ng + \sum_{k=0}^n \left[b^{n-k}(s) + \sum_{m=0}^{n-k} b^{n-k-m}(s) t_p(m) \right] T_{x-s}(k).$$

Ferner existiert

$$\begin{aligned} \lim_{n \rightarrow \infty} b^n(x) &= b(x) \\ &= L(x) + \sum_{u=0}^{x-s} L(x-u)m(u) + K - g \left[1 + M(x-s) \right]. \end{aligned}$$

Damit hat man eine Menge von linear unabhängigen Gleichungen mit den unbekannten Werten für $b(x)$ und g erhalten. Jedoch übersteigt die Zahl der Unbekannten die Zahl der Gleichungen um eine. Man kann somit einen der Werte für $b(x)$ gleich Null setzen, etwa $b(s) = 0$, und sodann die Werte für g und $b(x)$ bestimmen. Dabei erhält man für $b(x)$ nur relative Werte, die sich von den tatsächlichen durch eine additive Konstante unterscheiden. Aus der Gleichung $b(s) = 0$ erhält man dann den bekannten, vom Anfangsbestand unabhängigen, Ausdruck für die durchschnittlichen erwarteten Kosten g pro Periode eines stationären Modells

$$g = \frac{L(s) + \sum_{u=0}^D L(s-u)m(u) + K}{1 + M(D)}.$$

Durch Minimierung dieses Ausdrucks bezüglich der Parameter s und D erhält man ihre optimalen Werte für die vorgegebene (s, S) -Politik [2].

3. Das Zwei-Produkt/Ein-Lager Modell

Es werde jetzt der Fall betrachtet, daß in dem gleichen Lager gleichzeitig zwei Güter bevorratet werden. Die Unterscheidung der beiden Güter erfolge durch Induzierung der betreffenden Größen, etwa x_1 und x_2 , bzw. $L(y_1, y_2)$. Die Nachfrage nach den beiden Gütern werde in jeder Periode durch eine nicht-negative Zufallsvariable (U_1, U_2) bestimmt, deren Dichtefunktion $f(u_1, u_2)$ bekannt sei. Die Nachfrage aufeinanderfolgender Perioden werde wieder durch eine Folge unabhängiger und identisch verteilter Zufallsvariablen bestimmt. Die fixen Bestellkosten betragen jetzt $K(z_1, z_2)$, mit

$$K(z_1, z_2) = \begin{cases} K_1 & \text{für } z_1 > 0, z_2 = 0, \\ K_2 & \text{für } z_1 = 0, z_2 > 0, \\ K_{1,2} & \text{für } z_1 > 0, z_2 > 0, \\ 0 & \text{für } z_1 = 0, z_2 = 0, \end{cases}$$

sowie

$$\max(K_1, K_2) \leq K_{1,2} \leq K_1 + K_2.$$

Die Ein-Perioden Kosten für den Anfangsbestand (y_1, y_2) lauten jetzt

$$\begin{aligned} L(y_1, y_2) = & \int_0^\infty \int_0^{y_1} h_1(y_1 - u_1) f(u_1, u_2) du_1 du_2 + \int_0^{y_2} \int_0^\infty h_2(y_2 - u_2) f(u_1, u_2) du_1 du_2 \\ & + \int_0^\infty \int_{y_1}^\infty p_1(u_1 - y_1) f(u_1, u_2) du_1 du_2 + \int_{y_2}^\infty \int_0^\infty p_2(u_2 - y_2) f(u_1, u_2) du_1 du_2. \end{aligned}$$

Mit $C^n(x_1, x_2)$ wird der Erwartungswert der gesamten, entscheidungsabhängigen Kosten eines n -Perioden Problems bei einem Anfangsbestand von (x_1, x_2) Einheiten bezeichnet.

Um die Ergebnisse des Ein-Produkt/Ein-Lager Modells auf ein Zwei-Produkt/Ein-Lager Modell zu übertragen, muß zunächst die Form der Politik definiert werden, die für das Zwei-Produkt Modell befolgt werden soll.

In Anlehnung an die (s, S) -Politik des Ein-Produkt Modells und Überlegungen, die aus dem Ein-Perioden Problem [6] folgen, wird ein Punkt $P(S_1, S_2)$ als Schnittpunkt der beiden Geraden $x_1 = S_1$ und $x_2 = S_2$ in der zweidimensionalen euklidischen Ebene E^2 definiert. Durch eine Bestellung wird der Bestand jeweils bis zu diesem Punkt angehoben. Dadurch kann man sich für die weiteren Untersuchungen auf den Bereich

$$B := \{(x_1, x_2) \mid x_1 \leq S_1, x_2 \leq S_2\}$$

der zweidimensionalen Euklidischen Ebene E^2 beschränken. Dieser Bereich muß sodann in zwei disjunkte Teilbereiche B_0 und $B_{1,2}$ aufgespalten werden.

Befindet sich der Lagerbestand zu Beginn einer Periode in dem Bereich B_0 , dann wird keine Bestellung aufgegeben. Ist dagegen der Lagerbestand zu Beginn einer Periode in den Bereich $B_{1,2}$ abgesunken, dann werden beide Güter gleichzeitig bestellt, und der Bestand auf (S_1, S_2) angehoben. Dabei entstehen die fixen Bestellkosten $K_{1,2}$.

Es bleibt die Aufgabe, die beiden Bereiche B_0 und $B_{1,2}$ voneinander zu trennen. Dies kann auf verschiedene Arten erfolgen.

A. Politik I:

Bei dieser erfolgt die Trennung durch die beiden Geraden $x_1 = s_1$ und $x_2 = s_2$, mit dem Schnittpunkt $P(s_1, s_2)$, der im Inneren des Bereichs B liegt (siehe Abb. 1).

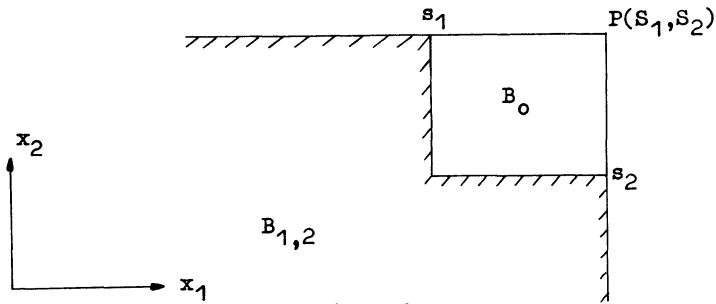


Abb. 1

Durch diese Trennung wird der Bereich B in die beiden Teilbereiche

$$B_0 : \{ (x_1, x_2) \mid s_1 < x_1, s_2 < x_2 \},$$

$$B_{1,2} : \{ (x_1, x_2) \mid x_1 \leq s_1, s_2 < x_2, \text{ oder } s_1 < x_1, x_2 \leq s_2, \text{ oder } x_1 \leq s_1, x_2 \leq s_2 \},$$

aufgeteilt.

Befindet sich der Lagerbestand zu Beginn einer Periode (x_1, x_2) im Bereich B_0 , dann wird keine Bestellung ausgelöst. Es wird angenommen, daß erstmals nach k Perioden der Lagerbestand in den Auslösebereich $B_{1,2}$ absinkt, sodaß nach $(k+1)$ Perioden eine Bestellung, d.h. eine Erneuerung stattfindet.

Es muß ein Ausdruck für die Wahrscheinlichkeit einer Bestellauslösung nach genau k Perioden gefunden werden. Die kumulierte Nachfrage während k aufeinanderfolgender Perioden nach den beiden Gütern werde durch

$$U_1^*(k) = U_1(1) + U_2(2) + \dots + U_k(k),$$

$$U_2^*(k) = U_2(1) + U_2(2) + \dots + U_k(k),$$

gegeben. Gilt

$$U_1^*(k) \leq x_1 - s_1 \quad \text{und} \quad U_2^*(k) \leq x_2 - s_2,$$

dann erfolgt auch nach k Perioden noch keine Bestellauslösung.

Da es nur die beiden Ereignisse "Bestellauslösung" und "Nicht-

bestellauslösung" gibt, läßt sich die Wahrscheinlichkeit für das Eintreffen des Ereignisses "Bestellauslösung" nach spätestens k Perioden als

$$1 - \text{Prob} \{ U_1^*(k) \leq x_1 - s_1, U_2^*(k) \leq x_2 - s_2 \}$$

formulieren. Will man die Wahrscheinlichkeit der Bestellauslösung nach genau k Perioden bestimmen, muß man diesen Ausdruck mit der Wahrscheinlichkeit multiplizieren, daß nach $(k-1)$ Perioden noch keine Bestellauslösung erfolgt war. Man erhält

$$\begin{aligned} & [1 - \text{Prob} \{ U_1^*(k) \leq x_1 - s_1, U_2^*(k) \leq x_2 - s_2 \}] \text{Prob} \{ U_1^*(k-1) \leq x_1 - s_1, U_2^*(k-1) \leq x_2 - s_2 \} \\ & = \text{Prob} \{ U_1^*(k-1) \leq x_1 - s_1, U_2^*(k-1) \leq x_2 - s_2 \} - \\ & \quad - \text{Prob} \{ U_1^*(k) \leq x_1 - s_1, U_2^*(k) \leq x_2 - s_2 \mid U_1^*(k-1) \leq x_1 - s_1, U_2^*(k-1) \leq x_2 - s_2 \} \\ & = \text{Prob} \{ U_1^*(k-1) \leq x_1 - s_1, U_2^*(k-1) \leq x_2 - s_2 \} - \text{Prob} \{ U_1^*(k) \leq x_1 - s_1, U_2^*(k) \leq x_2 - s_2 \} \\ & = F^{(k-1)}(x_1 - s_1, x_2 - s_2) - F^{(k)}(x_1 - s_1, x_2 - s_2). \end{aligned}$$

Die Wahrscheinlichkeit für das Eintreffen dieses Ereignisses nach genau k Perioden werde mit dem Ausdruck

$$T_{x_1 - s_1, x_2 - s_2}^{(k)}$$

bezeichnet. Setzt man $T_{D_1, D_2}^{(1)}(k) = T_{D_1, D_2}(k)$ für $k \geq 0$, wobei $D_i = s_i - s_i$ ($i = 1, 2$) gesetzt wurde, dann gilt

$$T_{D_1, D_2}^{(n)}(k) = \sum_{u=0}^k T_{D_1, D_2}^{(n-1)}(u) T_{D_1, D_2}(k-u)$$

für $k \geq 0$ und $n \geq 2$. Man definiert die Größe

$$t_{D_1, D_2}(k) = \sum_{n=1}^{\infty} T_{D_1, D_2}^{(n)}(k)$$

für $k \geq 0$, mit $t_{D_1, D_2}(0) = 0$.

Für $(x_1, x_2) \geq (s_1, s_2)$ und $n \geq 1$ gilt

$$\sum_{k=0}^n T_{x_1 - s_1, x_2 - s_2}^{(k)} = 1 - F^{(n)}(x_1 - s_1, x_2 - s_2),$$

$$\sum_{k=0}^n k T_{x_1-s_1, x_2-s_2}^{(k)} = 1 + \sum_{k=1}^{n-1} F^{(k)}(x_1-s_1, x_2-s_2) - n F^{(n)}(x_1-s_1, x_2-s_2).$$

Wegen

$$\lim_{n \rightarrow \infty} F^{(n)}(x_1, x_2) = 0, \quad \text{und} \\ \lim_{n \rightarrow \infty} \sum_{u=1}^n F^{(u)}(x_1, x_2) = M(x_1, x_2)$$

folgt

$$\sum_{k=0}^{\infty} T_{x_1-s_1, x_2-s_2}^{(k)} = 1, \quad \text{und} \\ \sum_{k=0}^{\infty} k T_{x_1-s_1, x_2-s_2}^{(k)} = 1 + M(x_1-s_1, x_2-s_2).$$

Aus der Konstruktion der vorgegebenen Politik ergibt sich für den Bereich $B_{1,2}$ die Kostenfunktion

$$B_{1,2}: \quad C^n(x_1, x_2) = K'_{1,2} + C^n(s_1, s_2) \quad \text{für } n \geq 1.$$

Für den Bereich B_0 erhält man für $n \geq 1$ die Kostenfunktion

$$C^n(x_1, x_2) = L(x_1, x_2) + \sum_{k=1}^{n-1} \sum_{u_1=0}^{x_1-s_1} \sum_{u_2=0}^{x_2-s_2} L(x_1-u_1, x_2-u_2) f^{(k)}(u_1, u_2) \\ + \sum_{k=1}^{n-1} \{K_{1,2} + C^{n-k}(s_1, s_2)\} T_{x_1-s_1, x_2-s_2}^{(k)}.$$

Diese Kosten werden wieder in Durchschnittskosten g pro Periode und ein Restglied $g^n(x_1, x_2)$

$$C^n(x_1, x_2) = ng + g^n(x_1, x_2) \quad \text{für } n \geq 1$$

zerlegt. Man erhält nach einigen Umformungen

$$g^n(x_1, x_2) = L(x_1, x_2) + \sum_{k=1}^{n-1} \sum_{u_1=0}^{x_1-s_1} \sum_{u_2=0}^{x_2-s_2} L(x_1-u_1, x_2-u_2) f^{(k)}(u_1, u_2) \\ + (K_{1,2} + ng) \sum_{k=1}^{n-1} T_{x_1-s_1, x_2-s_2}^{(k)} - g \sum_{k=1}^{n-1} k T_{x_1-s_1, x_2-s_2}^{(k)} \\ + \sum_{k=1}^{n-1} g^{n-k}(s_1, s_2) T_{x_1-s_1, x_2-s_2}^{(k)} - ng \\ = b^n(x_1, x_2) + \sum_{k=1}^{n-1} g^{n-k}(s_1, s_2) T_{x_1-s_1, x_2-s_2}^{(k)},$$

mit

$$b^n(x_1, x_2) = L(x_1, x_2) + \sum_{k=1}^{n-1} \sum_{u_1=0}^{x_1-s_1} \sum_{u_2=0}^{x_2-s_2} L(x_1-u_1, x_2-u_2) f^{(k)}(u_1, u_2) \\ + K_{1,2} \left[1 - F^{(n-1)}(x_1-s_1, x_2-s_2) \right] \\ - g \left[1 + \sum_{k=1}^{n-1} F^{(k)}(x_1-s_1, x_2-s_2) \right], \text{ und}$$

$$b^n(x_1, x_2) = 0,$$

für $(x_1, x_2) \geq (s_1, s_2)$. Speziell erhält man für $(x_1, x_2) = (s_1, s_2)$ die Erneuerungsgleichung

$$g^n(s_1, s_2) = b^n(s_1, s_2) + \sum_{k=0}^n g^{n-k}(s_1, s_2) t_{D_1, D_2}(k), \quad n \geq 0$$

mit der Lösung

$$g^n(s_1, s_2) = b^n(s_1, s_2) + \sum_{k=0}^n b^{n-k}(s_1, s_2) t_{D_1, D_2}(k).$$

Durch Kombination erhält man für die einzelnen Bereiche

a. $(x_1, x_2) \in B_{1,2}$:

$$G^n(x_1, x_2) = K_{1,2} + ng + b^n(s_1, s_2) + \sum_{k=0}^n b^{n-k}(s_1, s_2) t_{D_1, D_2}(k),$$

b. $(x_1, x_2) \in B_0$:

$$G^n(x_1, x_2) = b^n(x_1, x_2) + ng \\ + \sum_{k=0}^n \left[b^{n-k}(s_1, s_2) + \sum_{m=0}^{n-k} b^{n-k-m}(s_1, s_2) t_{D_1, D_2}(m) \right] T_{x_1-s_1, x_2-s_2}(k).$$

Ferner gilt

$$\lim_{n \rightarrow \infty} b^n(x_1, x_2) = b(x_1, x_2) \\ = L(x_1, x_2) + \sum_{u_1=0}^{x_1-s_1} \sum_{u_2=0}^{x_2-s_2} L(x_1-u_1, x_2-u_2) m(u_1, u_2) \\ + K_{1,2} - g \left[1 + M(x_1-s_1, x_2-s_2) \right],$$

mit

$$M(x_1, x_2) = \sum_{n=1}^{\infty} F^{(n)}(x_1, x_2).$$

Betrachtet man speziell die Beziehung für $b(S_1, S_2)$ und setzt diese gleich Null, so erhält man einen vom Anfangsbestand unabhängigen Ausdruck für die durchschnittlichen erwarteten Kosten g pro Periode eines stationären Modells

$$g = \frac{L(S_1, S_2) + \sum_{u_1=0}^{D_1} \sum_{u_2=0}^{D_2} L(S_1 - u_1, S_2 - u_2) m(u_1, u_2) + K_{1,2}}{1 + M(D_1, D_2)}.$$

Die optimalen Werte für die Parameter dieser Politik erhält man, indem man die partiellen Ableitungen von g nach s_i und D_i ($i=1,2$) gleich Null setzt.

B. Politik II:

Man kann weitere Abgrenzungen der Bereiche B_0 und $B_{1,2}$ voneinander betrachten. Jetzt erfolge die Trennung der beiden Bereiche durch eine Treppenfunktion. Diese werde von den Geraden

$$x_1 = s_1, x_1 = c_1, x_1 = S_1, \text{ mit } s_1 < c_1 < S_1, \text{ und}$$

$$x_2 = s_2, x_2 = c_2, x_2 = S_2, \text{ mit } s_2 < c_2 < S_2$$

gebildet (siehe Abb. 2).

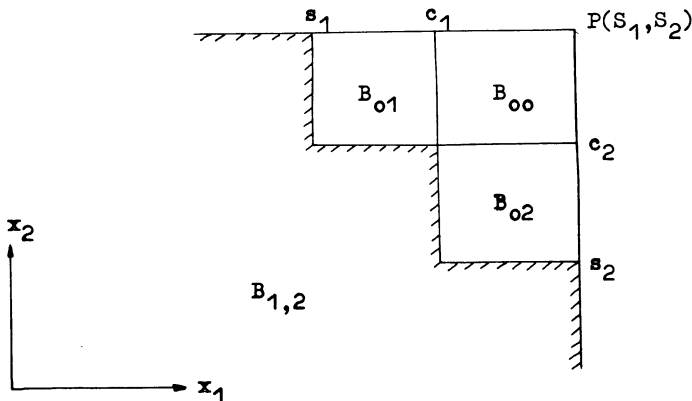


Abb. 2

Der Bereich B_0 setzt sich aus den Teilbereichen B_{00} , B_{01} und B_{02} wie folgt zusammen:

$$B_0 = \{B_{00}, B_{01}, B_{02}\},$$

mit

$$B_{00}: \{(x_1, x_2) \mid c_1 < x_1 \leq s_1, c_2 < x_2 \leq s_2\},$$

$$B_{01}: \{(x_1, x_2) \mid s_1 < x_1 \leq c_1, c_2 < x_2 \leq s_2\},$$

$$B_{02}: \{(x_1, x_2) \mid c_1 < x_1 \leq s_1, s_2 < x_2 \leq c_2\}.$$

Der Bereich $B_{1,2}$ wird wie folgt definiert

$$B_{1,2}: \{(x_1, x_2) \mid x_1 \leq c_1, x_2 \leq c_2, \text{ oder} \\ x_1 \leq s_1, c_2 < x_2 \leq s_2, \text{ oder} \\ c_1 < x_1 \leq s_1, x_2 \leq s_2\}.$$

Um die Kostenfunktion aufstellen und lösen zu können, benötigt man einen Ausdruck für die Wahrscheinlichkeit einer Bestellauslösung nach genau k Perioden. Für einen Anfangsbestand (x_1, x_2) aus den Bereichen B_{01} und B_{02} folgt dieser Ausdruck unmittelbar aus den bei der Politik I angestellten Überlegungen:

$$(x_1, x_2) \in B_{01}:$$

$$T_{x_1-s_1, x_2-c_2}^{(k)} = F^{(k-1)}(x_1-s_1, x_2-c_2) - F^{(k)}(x_1-s_1, x_2-c_2),$$

$$(x_1, x_2) \in B_{02}:$$

$$T_{x_1-c_1, x_2-s_2}^{(k)} = F^{(k-1)}(x_1-c_1, x_2-s_2) - F^{(k)}(x_1-c_1, x_2-s_2).$$

$$(x_1, x_2) \in B_{00}:$$

Mit der Wahrscheinlichkeit

$$F^{(k-1)}(x_1-s_1, x_2-c_2) - F^{(k)}(x_1-s_1, x_2-c_2)$$

erfolgt nach genau k Perioden der Übergang aus dem Bereich $B_{00} \cup B_{01}$ in den Bereich $B_{1,2} \cup B_{02}$. Da nur eine Nachfrage nach positiven Gütern definiert ist, bedeutet dies, daß durch den obigen Ausdruck die folgenden Übergänge erfaßt werden

$$B_{00} \rightarrow B_{1,2}, B_{00} \rightarrow B_{02}, B_{01} \rightarrow B_{1,2}.$$

Ebenso erfolgt mit der Wahrscheinlichkeit

$$F^{(k-1)}(x_1-c_1, x_2-s_2) - F^{(k)}(x_1-c_1, x_2-s_2)$$

nach genau k Perioden der Übergang aus dem Bereich $B_{00} \cup B_{02}$ in den Bereich $B_{1,2} \cup B_{01}$. Dies bedeutet, daß durch diesen Ausdruck die folgenden Übergänge erfaßt werden

$$B_{00} \rightarrow B_{01}, B_{00} \rightarrow B_{1,2}, B_{02} \rightarrow B_{1,2}.$$

Mit der Wahrscheinlichkeit

$$F^{(k-1)}(x_1-c_1, x_2-c_2) - F^{(k)}(x_1-c_1, x_2-c_2)$$

erfolgt nach genau k Perioden der Übergang aus dem Bereich B_{00} in den Bereich $B_{1,2} \cup B_{01} \cup B_{02}$. Mit diesem Ausdruck werden die folgenden Übergänge erfaßt

$$B_{00} \rightarrow B_{01}, B_{00} \rightarrow B_{02}, B_{00} \rightarrow B_{1,2}.$$

Bildet man nun den Ausdruck

$$\begin{aligned} & F^{(k-1)}(x_1-s_1, x_2-c_2) - F^{(k)}(x_1-s_1, x_2-c_2) \\ & + F^{(k-1)}(x_1-c_1, x_2-s_2) - F^{(k)}(x_1-c_1, x_2-s_2) \\ & - \left[F^{(k-1)}(x_1-c_1, x_2-c_2) - F^{(k)}(x_1-c_1, x_2-c_2) \right], \end{aligned}$$

so werden damit genau die Übergänge

$$B_{00} \rightarrow B_{1,2}, B_{01} \rightarrow B_{1,2}, B_{02} \rightarrow B_{1,2}$$

in der k -ten Periode erfaßt. Damit wurde der gesuchte Ausdruck für die Wahrscheinlichkeit der Bestellauslösung nach genau k Perioden gefunden. Er werde zur Abkürzung mit

$$(T_{x_1-s_1, x_2-c_2} + T_{x_1-c_1, x_2-s_2} - T_{x_1-c_1, x_2-c_2})^{(k)}$$

bezeichnet.

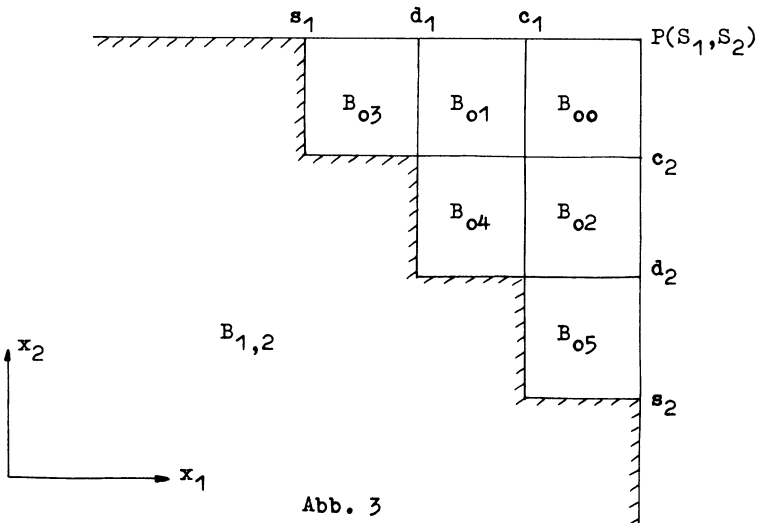


Abb. 3

Der Bereich B_0 setzt sich jetzt aus den Teilbereichen B_{00} , B_{01} , B_{02} , B_{03} , B_{04} und B_{05} zusammen.

Befindet sich der Lagerbestand zu Beginn einer Periode (x_1, x_2) in den Bereichen B_{01} , B_{02} , B_{03} , B_{04} oder B_{05} , so ergibt sich die Wahrscheinlichkeit einer Bestellauslösung nach genau k Perioden unmittelbar aus den Überlegungen, die bereits bei den Politiken I und II angestellt wurden.

Es bleibt der Fall zu untersuchen, daß der Lagerbestand zu Beginn einer Periode (x_1, x_2) im Bereich B_{00} liegt. Bildet man den Ausdruck

$$\begin{aligned} & (T_{x_1-s_1, x_2-c_2} + T_{x_1-d_1, x_2-d_2} + T_{x_1-c_1, x_2-s_2} \\ & - T_{x_1-d_1, x_2-c_2} - T_{x_1-c_1, x_2-d_2})(k) \end{aligned}$$

so zeigen analoge Betrachtungen, wie sie bereits bei der Politik II angestellt wurden, daß damit genau die Übergänge

Man setzt $D_i = S_i - s_i$ und $E_i = S_i - c_i$ für $i = 1, 2$ und erhält nach Überlegungen, die völlig analog zu den bei der Politik I angestellten verlaufen, für die vom Anfangsbestand unabhängigen durchschnittlichen erwarteten Kosten g pro Periode eines stationären Modells den Ausdruck

$$g = \frac{b_{00}(S_1, S_2)}{1 + M(D_1, E_2) + M(E_1, D_2) - M(E_1, E_2)},$$

mit

$$\begin{aligned} b_{00}(S_1, S_2) &= K_{1,2} + L(S_1, S_2) \\ &+ \sum_{u_1=0}^{D_1} \sum_{u_2=0}^{E_2} L(S_1 - u_1, S_2 - u_2) m(u_1, u_2) \\ &+ \sum_{u_1=0}^{E_1} \sum_{u_2=0}^{D_2} L(S_1 - u_1, S_2 - u_2) m(u_1, u_2) \\ &+ \sum_{u_1=0}^{E_1} \sum_{u_2=0}^{E_2} L(S_1 - u_1, S_2 - u_2) m(u_1, u_2). \end{aligned}$$

Wieder erhält man die optimalen Werte für die Parameter dieser Politik, indem die partiellen Ableitungen von g nach s_i , c_i , D_i und E_i ($i=1, 2$) gleich Null setzt.

C. Politik III:

Die Trennung der beiden Bereiche B_0 und $B_{1,2}$ durch eine Treppenfunktion kann weiter verfeinert werden. Sie erfolge jetzt durch die Geraden

$$x_1 = s_1, x_1 = d_1, x_1 = c_1 \text{ und } x_1 = S_1, \text{ mit } s_1 < d_1 < c_1 < S_1,$$

$$x_2 = s_2, x_2 = d_2, x_2 = c_2 \text{ und } x_2 = S_2, \text{ mit } s_2 < d_2 < c_2 < S_2$$

(siehe Abb. 3).

$$B_{00} \rightarrow B_{1,2}, B_{01} \rightarrow B_{1,2}, B_{02} \rightarrow B_{1,2}, B_{03} \rightarrow B_{1,2}, \\ B_{04} \rightarrow B_{1,2}, B_{05} \rightarrow B_{1,2}$$

in der k -ten Periode erfaßt werden. Damit hat man den gesuchten Ausdruck für die Wahrscheinlichkeit des Übergangs nach genau k Perioden aus dem Bereich B_{00} in den Bestellbereich $B_{1,2}$ gefunden. Zur Abkürzung setzt man $D_1 = S_1 - s_1$, $E_1 = S_1 - c_1$ und $F_1 = S_1 - d_1$, für $i = 1, 2$, und erhält nach einigen Rechnungen für die vom Anfangsbestand unabhängigen durchschnittlichen erwarteten Kosten g pro Periode eines stationären Modells den Ausdruck

$$g = \frac{b_{00}(S_1, S_2)}{1 + M(D_1, E_2) + M(F_1, F_2) + M(E_1, D_2) - M(F_1, E_2) - M(E_1, F_2)},$$

mit

$$b_{00}(S_1, S_2) = K_{1,2} + L(S_1, S_2) \\ + \left[\sum_{u_1=0}^{D_1} \sum_{u_2=0}^{E_2} + \sum_{u_1=0}^{F_1} \sum_{u_2=0}^{F_2} + \sum_{u_1=0}^{E_1} \sum_{u_2=0}^{D_2} \right. \\ \left. - \sum_{u_1=0}^{F_1} \sum_{u_2=0}^{E_2} - \sum_{u_1=0}^{E_1} \sum_{u_2=0}^{F_2} \right] L(S_1 - u_1, S_2 - u_2) m(u_1, u_2).$$

Wieder erhält man die optimalen Parameterwerte mit den bekannten Methoden der Differentialrechnung.

4. Literatur:

- [1] Feller, W., "An Introduction to Probability Theory and its Applications, Volume II", John Wiley, New York, 1966
- [2] Hochstädter, D., "Stochastische Lagerhaltungsmodelle", Springer-Verlag, Berlin 1969
- [3] Hochstädter, D., "Neuere Entwicklungen der stochastischen Lagerhaltungstheorie", Industrielle Organisation, Band 39, no. 12, 1970, pp 507-513, wieder abgedruckt in: "Unternehmensforschung Heute", Springer-Verlag, Berlin 1971, pp 30-51
- [4] Hochstädter, D., "The Stationary Solution of Multi-Product Inventory Models (The discounted cost criterion)", paper presented at the "International Conference on Inventory Control and Water Storage Problems", at Győr (Hungary), September 13-17, 1971, to be published
- [5] Rieß, G. "Vergleichende Untersuchung zur Bestimmung der optimalen Parameter von (s,S)-Politiken", Diplomarbeit TU München, Abteilung für Angewandte Mathematik, 1971
- [6] Sivazlian, B.D., "A Production and Inventory Control Model for a Multi-Commodity System with Probabilistic Demand and Interacting Set-Up Cost", Technical Report No. 1, Department of Industrial and Systems Engineering, The University of Florida, 1967
- [7] Veinott, A.F.jr and H.M. Wagner, "Computing Optimal (s,S) Inventory Policies", Management Science, vol. 11, no. 5, 1965, pp 525-552

Sinusoidal Functions for Inventory Control Models

by M. A. Tawadros, South Bend

Introduction:

Inventory models are mathematical formulations developed to find an optimal inventory policy that will permit rational decisions concerning the order point and the order quantity. The order point is a function of the lead time, while the order quantity is a function of the expected demand. Nonetheless, the main uncontrollable variable is the expected demand. If demand is known with certainty and the lead time is fixed, then there will be no risk to face and decisions made to optimize the inventory control system will be perfect under these highly hypothetical assumptions. On the other hand, if demand is known with certainty while the lead time may vary, then the only risk involved will be the depletion of the safety stock, if, and only if, the lead time is longer than expected. However, as long as the safety stock is not entirely depleted, there is no real threat of considering the cost of shortage to be included in the inventory model.

A realistic situation exists when either fixed or variable lead time occurs with uncertain demand. Expected demand is largely independent of

The author wishes to express his deepest appreciation to his colleagues: Professors G. Harriman, G. Wing, L. Waltz, and E. Albert, Business and Economics Division, IUSB, for their encouragement and their helpful suggestions.

the firm's control.¹ As a result, all inventory control models have been based on the assumption that demand is a random variable with a known probability distribution. [3] [5] [6] [8]

Even though, the foregoing assumption is convenient to consider in the construction of any mathematical inventory control model, and in fact, has served to derive optimal inventory policies, nonetheless, a reasonable accurate demand forecast [7] is more powerful than any probabilistic demand.

Inventory as well as demand patterns change with time and periodic fluctuations exist in both. It is appropriate to use a periodic function² to predict the oscillatory characteristics of the inventory or the demand functions. Therefore, the purpose of this paper is to find the best fit of regression estimate of a sinusoidal function to be applied for the prediction of demand or inventory.

Model:

Periodic variations of demand are due to the components of time series. [4] Inventory has more periodic variations than demand or sales because not only the demand varies but the lead time may vary as well. The most significant characteristic of the periodic fluctuations of either demand or inventory is the fluctuations are not of the same length or amplitude.

Periodic fluctuations suggest the use of a trigonometric function. These functions are periodic but they have regular oscillations of 2π in length.

The trigonometric function can be extended to show a great variety of regular oscillations and is called the sinusoidal function. Trigonometric

¹Promotion, advertising, and price variations may influence demand.

²The pattern of demand has been described by a time series model as follows:
 $X_t = a_1 + a_2 t + a_3 \sin 30t^\circ + a_4 \cos 30t^\circ + a_5 \sin 60t^\circ + a_6 \cos 60t^\circ$
 For more details about this model, see Robert Brown, Decision Rules for Inventory Management, New York: Holt, Rinehart and Winston, 1967, pp. 118-24.

functions as well as sinusoidal functions have three characteristics: amplitude, period, and phase. These features are represented by the three parameters: A, W, and E respectively, as shown below in equations (1) and (2):

$$Y = A \sin (WX+E) \quad (1)$$

$$Y = A \cos (WX+E) \quad (2)$$

If the amplitude of the trigonometric function varies with X according to some function A(X); such that $A(X) = Ae^{\alpha x}$; then the function may stay constant, converge or diverge according to the magnitude of α . Accordingly, a regular, damped, or anti-damped function can be derived from equation (3) or (4):

$$Y = Ae^{\alpha x} \sin (WX+E) \quad (3)$$

$$Y = Ae^{\alpha x} \cos (WX+E) \quad (4)$$

The only difference between the two sets of equations (1), (2) and (3), (4) is the change of the amplitude or the effect of damping or anti-damping, but the upswings and downswings are of the same cosine wave.

Oscillations of inventory and demand functions as well as most economic variables are not of the same length or regular cosine shape. [1] This means that a single sinusoidal function (either equation (3) or (4)) will not be appropriate to fit the periodic variations that exist in inventory or demand functions. Therefore, a sum of sinusoidal functions with different harmonics, as shown in equation (5) is applied to reflect the change in the length of the cycles.

$$Y = \sum_{i=1}^n A \cos (WX+E) \quad (5)$$

or

$$Y = \sum_{i=1}^n A (\cos WX \cos E - \sin WX \sin E)$$

$$Y = \sum (a_i \cos WX + b_i \sin WX)$$

To estimate the parameters of equation (5), the least square method [2] is used and the estimated periodic equation becomes:

$$\hat{Y} = A_0 + \sum_{i=1}^n (a_i \cos c + b_i \sin c) \quad (6)$$

where: $a = A \cos E$ $b = -A \sin E$
 $c = WX = 2\pi/T$ $T = \text{the number of time values}$

It is important to mention that the number of sinusoidal oscillations or harmonics in the estimated equation (6) should be sufficient to predict the ups and downs of the periodic function over a reasonable period of time.

Method of Analysis:

Equation (6), the estimated periodic equation, can be used to find the parameters of the best fit for time series data of inventory or demand.

If inventory records are available, then such data will be used to calculate the parameters of the inventory function and to forecast its periodic fluctuations.³ An accurate forecast of the inventory function will help in the construction of an effective computerized inventory control system.⁴

On the other hand, if inventory records are not available, the demand^{*} data may be used to forecast the demand periodic function. Given an accurate demand forecast, an effective inventory control system can be programmed provided that information of the lead time is available.

In this investigation, forecast of the demand periodic function is considered rather than the inventory function. The same method of analysis can be applied if inventory data are available.

³Assuming that the past inventory performance was satisfactory to maintain a proper inventory level.

⁴See Appendix I.

A periodic equation based on sinusoidal functions as defined in equation (6) with $i=1$ through 9 harmonics has been applied to seasonally unadjusted time series sales data of the following activities: estimated real sales of retailed nondurable goods stores for the U.S., estimated real sales of retailed durable goods stores for the U.S.,⁵ and sales of natural gas for both industrial and commercial purposes in the city of South Bend, Indiana.⁶

Monthly data have been used for each estimated equation for different lengths of time: eight, five, and three-year periods. However, it is found that monthly data for a three-year period yields the best prediction, therefore, monthly data for 1966, 1967 and 1968 are the observed values used in each model.

For each model, nine estimated equations of different harmonics have been estimated and tested to select the best according to the following criterion:

1. Comparing the observed, predicted, and percentages of residuals to observed data.
2. The overall fit of the derived equations should be satisfactory as measured by the coefficient of determination (R^2).

Findings:

Parameter estimates and R^2 were calculated for 36 equations, nine for each periodic regression model with harmonics ranging from one through nine, using monthly data for the three-year period 1966-1968. Each estimated equation has been analyzed and tested and the best four periodic regression

⁵U.S. Department of Commerce: Survey of Current Business, December issues from 1967 to 1969.

⁶Data compiled by the Bureau of Business and Economic Research, Business and Economic Division, IUSB.

equations are presented in the following order:

Equation 1 - represents estimated real sales of retailed nondurable goods stores for the U.S.

Equation 2 - represents estimated real sales of retailed durable goods stores for the U.S.

Equation 3 - shows sales of natural gas for industrial purposes, South Bend, Indiana.

Equation 4 - shows sales of natural gas for commercial purposes, South Bend, Indiana.

Equations 1 and 2 are estimates for demand of the discrete type, while Equations 3 and 4 are for the continuous type.

Equation 1:

$$\begin{aligned}\hat{Y}_t = & 15691.78 + 3.08 \cos (ct) - 404.90 \sin (ct) \\ & - 107.26 \cos (2ct) - 194.94 \sin (2ct) \\ & + 53.64 \cos (3ct) - 749.50 \sin (3ct) \\ & - 23.61 \cos (4ct) - 90.89 \sin (4ct) \\ & - 53.92 \cos (5ct) - 64.97 \sin (5ct) \\ & - 11.39 \cos (6ct) - 1202.81 \sin (6ct) \\ & - 49.59 \cos (7ct) - 63.27 \sin (7ct) \\ & - 64.25 \cos (8ct) - 44.38 \sin (8ct) \\ & - 326.39 \cos (9ct) - 1120.61 \sin (9ct)\end{aligned}$$

$$R^2 = .66$$

The magnitude of the coefficient of determination R^2 shows that 66% of the total variation of the dependent variable has been explained by the regression Equation 1. A graphical presentation of the observed and predicted values of this equation are shown in Figure 1 in which the ups and downs of the real sales can be determined very easily and can be used to reflect the variation in the stock on hand.

Equation 2:

$$\begin{aligned}
\hat{Y}_t = & 8148.00 + 193.23 \cos (ct) + 292.28 \sin (ct) \\
& - 170.07 \cos (2ct) + 139.87 \sin (2ct) \\
& - 439.73 \cos (3ct) - 30.91 \sin (3ct) \\
& - 19.90 \cos (4ct) + 85.25 \sin (4ct) \\
& + 65.72 \cos (5ct) + 44.16 \sin (5ct) \\
& - 262.20 \cos (6ct) - 465.78 \sin (6ct) \\
& + 4.79 \cos (7ct) - 47.02 \sin (7ct) \\
& + 67.47 \cos (8ct) + 105.87 \sin (8ct) \\
& - 273.50 \cos (9ct) - 111.39 \sin (9ct)
\end{aligned}$$

$$R^2 = .80$$

The R^2 value of .80 indicates that the equation fits the data well. Moreover, Figure 2 shows that the predicted curve reflected the fluctuations of the observed data.

Equation 3:

$$\begin{aligned}
\hat{Y}_t = & 1278.50 - 13.53 \cos (ct) - 92.69 \sin (ct) \\
& - 44.12 \cos (2ct) + 5.98 \sin (2ct) \\
& + 342.93 \cos (3ct) + 49.92 \sin (3ct) \\
& - 19.29 \cos (4ct) - 65.97 \sin (4ct) \\
& - 26.06 \cos (5ct) - 39.96 \sin (5ct) \\
& - 19.83 \cos (6ct) - 24.54 \sin (6ct) \\
& - 20.27 \cos (7ct) - 51.36 \sin (7ct) \\
& - 14.95 \cos (8ct) - 22.58 \sin (8ct) \\
& + 26.50 \cos (9ct) + 4.61 \sin (9ct)
\end{aligned}$$

$$R^2 = .94$$

Equation 3 fits the data very well as indicated by the magnitude of R^2 . Moreover, the predicted values fluctuate very closely with the observed data as shown in Figure 3.

Equation 4:

$$\begin{aligned}
\hat{Y}_t = & 1848.94 - 109.29 \cos (ct) - 153.87 \sin (ct) \\
& - 95.96 \cos (2ct) + 35.86 \sin (2ct) \\
& + 1168.19 \cos (3ct) + 537.85 \sin (3ct) \\
& - 37.20 \cos (4ct) - 157.88 \sin (4ct) \\
& - 34.63 \cos (5ct) - 1.74 \sin (5ct) \\
& + 87.69 \cos (6ct) + 18.14 \sin (6ct) \\
& - 9.63 \cos (7ct) - 111.16 \sin (7ct) \\
& + 3.17 \cos (8ct) - 32.89 \sin (8ct) \\
& - 24.83 \cos (9ct) - 60.61 \sin (9ct)
\end{aligned}$$

$$R^2 = .98$$

The value of R^2 indicates that 98% of the variation of the demand has been accounted for by the regression equation 4. Additionally, the predicted values swing up and down very closely to the observed values as shown in Figure 4.

The results of the findings are satisfactory as shown by the magnitude of R^2 , and the graphical presentation of actual and predicted values.⁷ The satisfactory estimates of these four equations provide policy makers with a powerful tool that can be used to predict the demand.

Implication of the Findings for Demand Prediction:

As stated previously, the main goal of this paper is to find a model that can be used to predict with accuracy the expected demand in order to construct an effective computerized inventory control system. The question of the model's merit is now at hand. If the model's predictive ability is good, then it is reasonable to assume that the model is a reliable

⁷Graphical presentation of actual and predicted values are presented in Figures 1 through 4.

representation of the structure involved, and therefore, the model can explain the structure's behavior when confronted by a set of empirical observations that differ from those used to determine the coefficients of the model.

Prediction is often separated into two types: ex-ante and ex-post. Ex-ante prediction is true forecasting and depends on correctly guessing the values of the explanatory variables, which are, in most cases, unknown at the time the forecasting takes place. Ex-post prediction, on the other hand, is merely a convenient way by which information can be obtained regarding the degree to which an equation represents a true picture of economic behavior. For the purpose of policy recommendation, ex-ante prediction is the only practicable type. However, in this study ex-post prediction is used to assess the reliability of the model. If the model proves to be reliable with ex-post prediction, then it will be recommended for ex-ante prediction.

The estimated periodic equation (6) served to find coefficients of the regression equations 1 through 4, which in turn have been used to calculate predicted values of 1966, 1967 and 1968. However, management is more interested in ex-ante prediction; therefore, it is vital to use the coefficients of the estimated equations to predict values of a time period subsequent to the period of the sample (e.g., 1969 values).

The use of estimated coefficients for 1966 to 1968 to predict 1969 values has succeeded in reflecting the periodic fluctuations; however, it failed to account for the growth of demand and sales in 1969.⁸ Therefore, two major changes have been devised for the estimated equation (6), namely:

1. The multiplication of estimates by a growth rate coefficient β

⁸The use of the logarithm of sinusoidal functions is under investigation.

which has been calculated for the last year of the sample (1968), assuming that sales or demand develops according to an exponential function: $Y = A_0 \beta^t$.

2. Using different estimates of the constant A_0 which represents \bar{Y} , the average sales of the three years (1966, 1967 and 1968). Other values that A_0 has assumed are: the average of the last two years (1967 and 1968), and the average of the last year (1968).

Accordingly, the estimated equation used for ex-ante prediction becomes:

$$\hat{Y} = A_0 + \beta \sum_{i=1}^n (a_i \cos c + b_i \sin c) \quad (7)$$

Predicted sales of the four regression equations are presented in Tables 1 through 4. Each table has five columns, the first column shows the observed values of sales in 1969, while the other four columns have predicted values of sales calculated by the following criterion:

- a. \hat{Y} represents the estimates of equation (6).
- b. $\hat{Y}_1 = Y(1 + \beta)$, same as (a) except raising the curve by a constant rate of growth (β).
- c. \hat{Y}_2 , same as (b), but A_0 is the average sales of 1967 and 1968.
- d. \hat{Y}_3 , same as (b), but A_0 is the average sales of 1968 (the last year).

Additionally, R^2 , RMS^9 , and U-coefficient¹⁰ have been calculated for each predicted set of values to determine the best method of forecasting to be recommended for ex-ante prediction.

Tables 1 through 4 indicate that the second and the third criterion

$$^9RMS = \sqrt{\sum (P_i - A_i)^2 / N}$$

¹⁰ $U = \sqrt{\sum (P_i - A_i)^2 / \sum A_i^2}$ for more details about RMS and U-coefficient Henri Theil, Applied Economic Forecasting, Amsterdam, North Holland Publishing Co., 1966, Chapter 2.

of prediction, \hat{Y}_1 and \hat{Y}_2 , are the best as shown by the magnitudes of R^2 , RMS, and the U-coefficient.

Values of R^2 for \hat{Y}_1 and \hat{Y}_2 of all the four regression equations are very significant and range between .80 and .97.

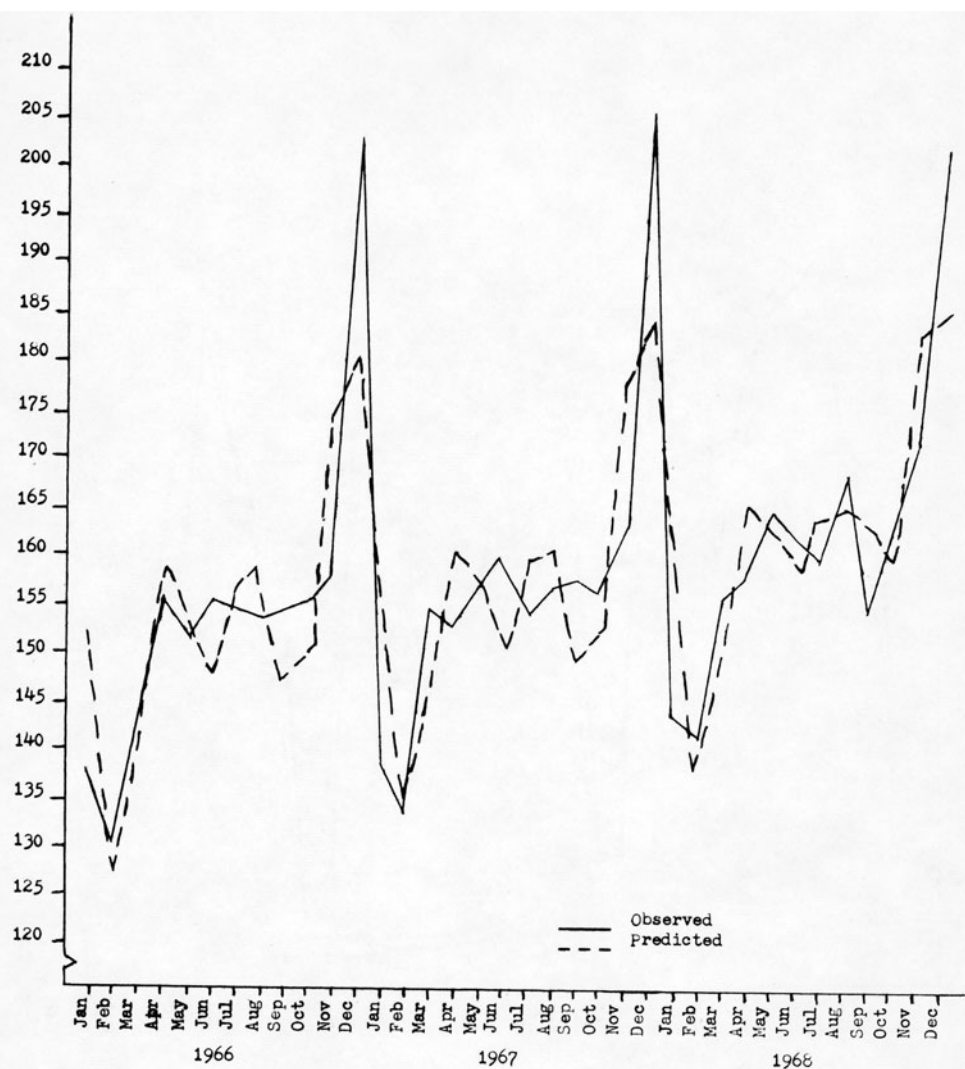
On the other hand, the root-mean-square prediction error, RMS, used to measure the seriousness of a given forecast error reveals that the percentage forecasting error per month for the four regression equations are 0.9%, 0.7%, 2.7%, and 8.0% respectively for \hat{Y}_1 and \hat{Y}_2 . These percentages show the insignificance of the forecast error for Equations 1, 2 and 3 which range from 0.7% to 2.7%. For Equation 4, even though the RMS prediction error is as high as 8.0%, however, it is within the acceptable range.

Additionally, U-coefficient values of 1.02 and 1.13 for Equations 1 and 2 respectively almost lead to the same RMS prediction error as naive no-change extrapolation, while the values of U-coefficient of .83 and .68 for Equations 3 and 4 respectively indicate that the RMS prediction error is 83% and 68% of that RMS error of the naive no-change prediction model. The U-coefficient values indicate that the model's predictive ability, in general, is better than that of its naive model.

Conclusions:

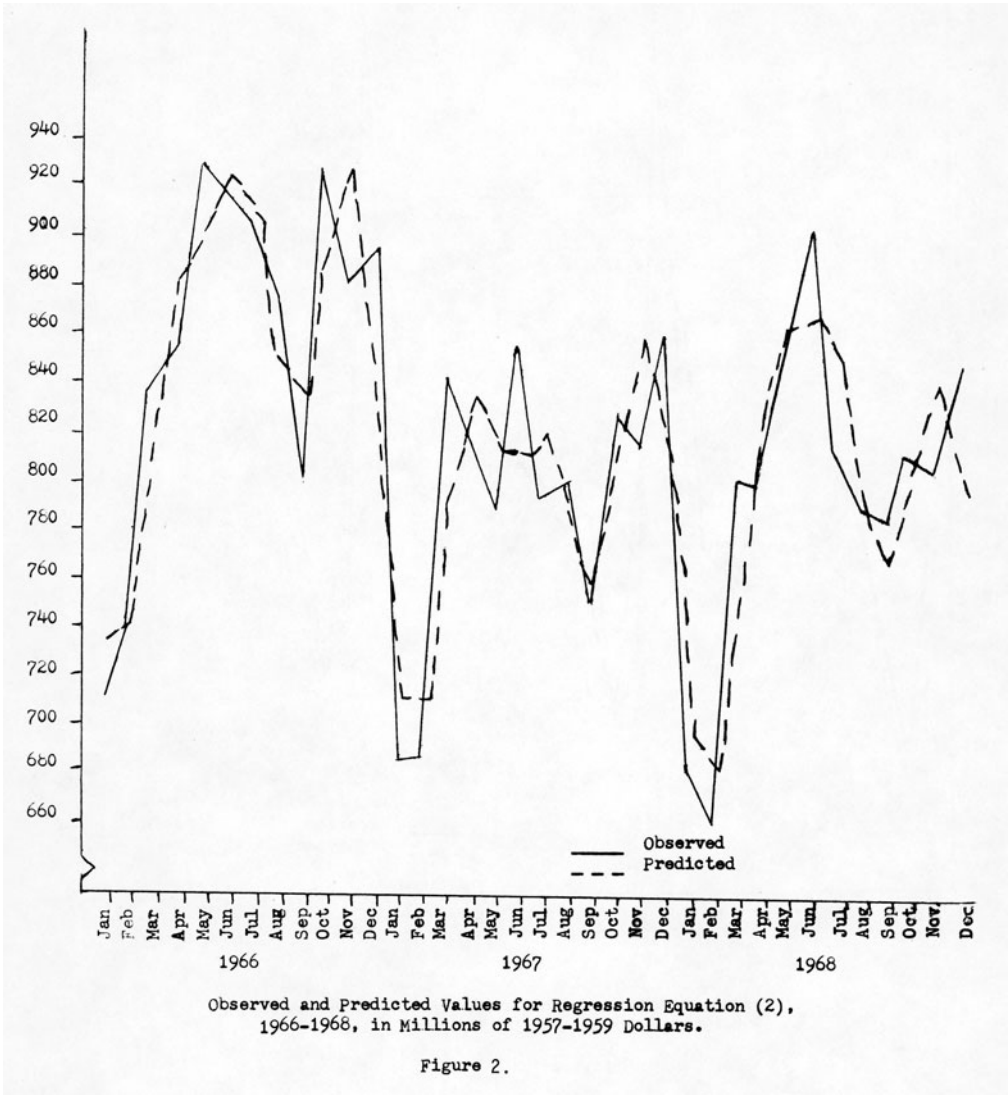
The results of the findings of this study are satisfactory and lead to the following conclusions:

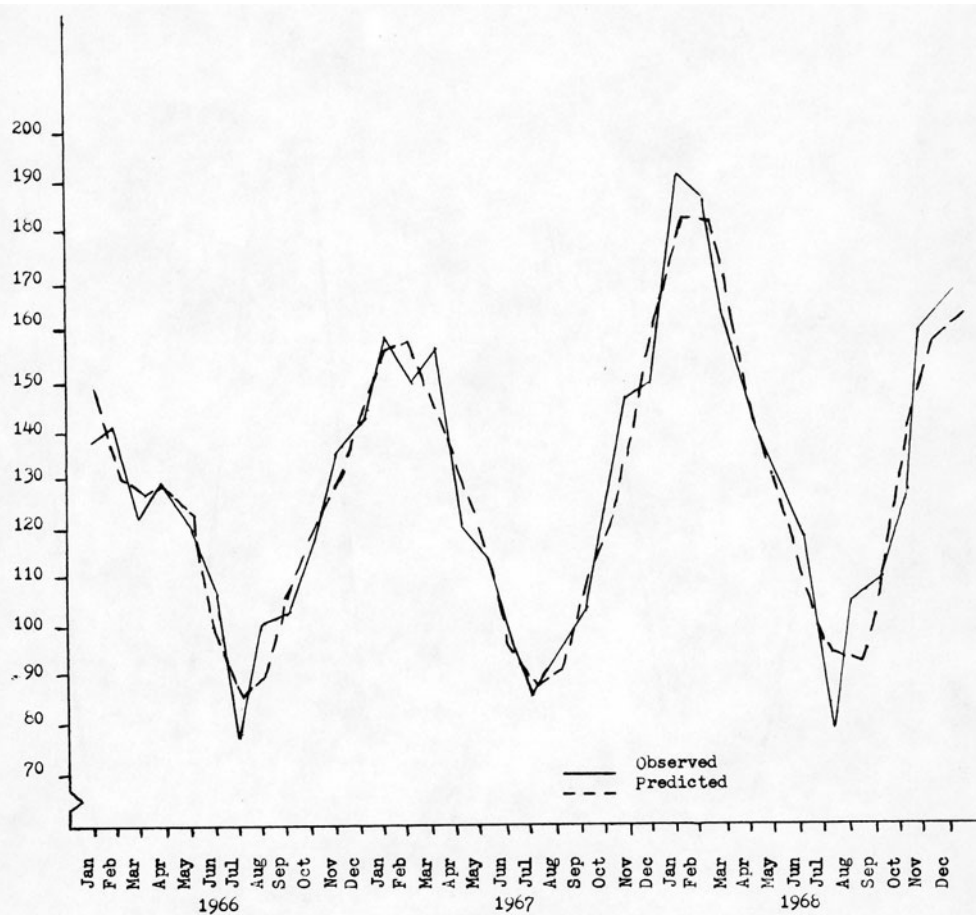
1. Sinusoidal functions proved to be very appropriate in predicting the oscillatory characteristics of demand functions and can provide management with reliable demand forecasts that can be considered as the cornerstone of an effective computerized inventory control system.
2. The same analysis of this investigation is applicable to forecast a fairly accurate inventory function that can be used to construct a programmed dependable inventory control system.



Observed and Predicted Values for Regression Equation (1),
1966-1968, in Millions of 1957-1959 Dollars.

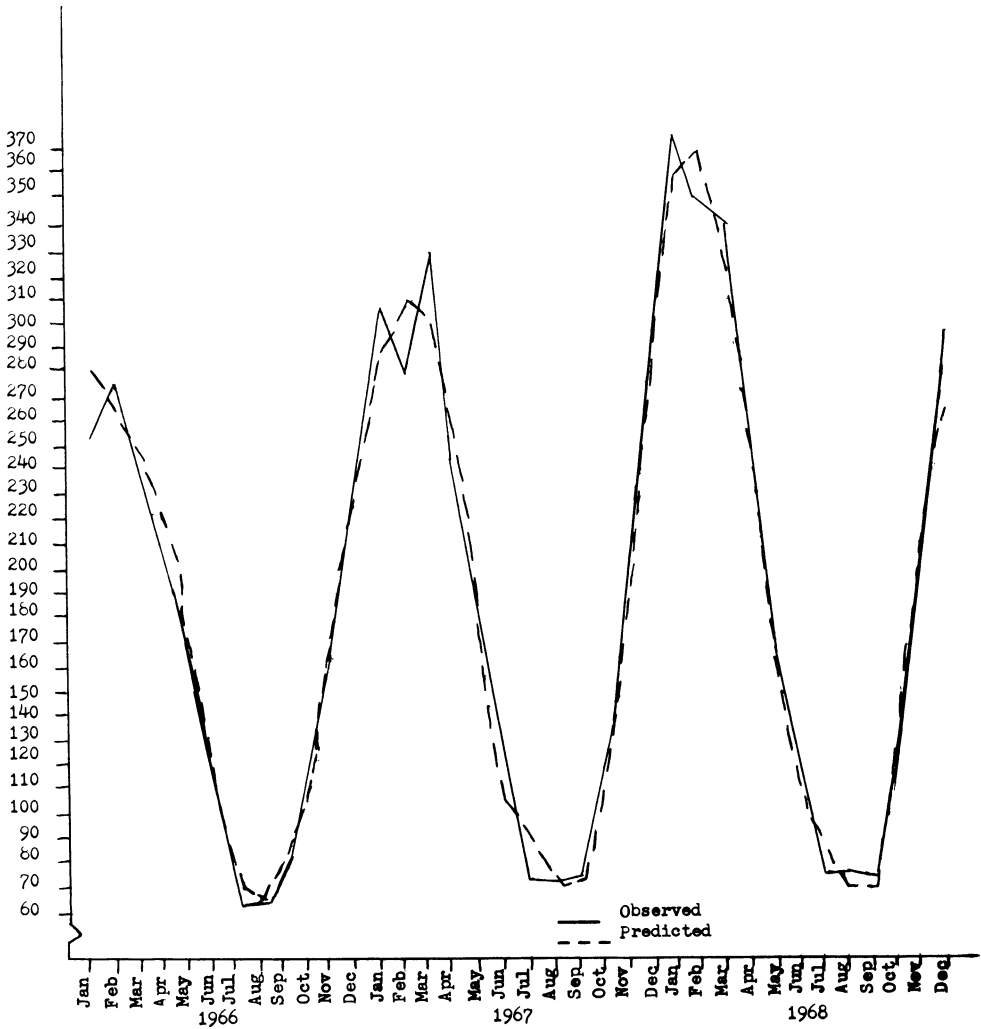
Figure 1.





Observed and Predicted Values for Regression Equation (3),
1966-1968, in 100,000 of Cubic Feet.

Figure 3.



Observed and Predicted Values for Regression Equation (4),
1966-1968, in 100,000 of Cubic Feet.

Figure 4.

Table 1
Observed, Predicted Values of 1969,
and Related Statistics for Regression Equation 1,
in Millions of 1957-1959 Dollars

	Observed		Predicted		
<u>1969</u>	Y	Y	Y ₁	Y ₂	Y ₃
Jan	14766	15112	15414	15553	15768
Feb	13680	12669	12922	13061	13275
Mar	15351	13911	14189	14328	14542
Apr	15542	15744	16059	16198	16412
May	16710	15365	15672	15811	16025
Jun	15647	14722	15016	15155	15369
Jul	15662	15523	15833	15972	16186
Aug	16355	15773	16088	16227	16441
Sep	15277	14646	14939	15078	15292
Oct	16328	14944	15243	15382	15596
Nov	16363	17353	17700	17839	18053
Dec	20476	18031	18392	18531	18745
R ²		.96	.97	.97	.97
RMS		.009	.009	.009	.009
U-coefficient		1.02	1.02	1.02	1.01

$\beta = .02$

Table 2

Observed, Predicted Values of 1969,
and Related Statistics for Regression Equation 2,
in Millions of 1957-1959 Dollars.

	Observed		Predicted		
<u>1969</u>	Y	Y	Y ₁	Y ₂	Y ₃
Jan	7674	7314	7416	7215	7225
Feb	7516	7424	7528	7327	7337
Mar	8301	8195	8310	8109	8119
Apr	8745	8783	8906	8705	8715
May	9717	9048	9175	8974	8984
Jun	9138	9248	9377	9176	9186
Jul	8486	9105	9232	9031	9041
Aug	7996	8513	8632	8431	8441
Sep	8391	8306	8422	8221	8231
Oct	8883	8931	9056	8855	8865
Nov	8024	9283	9413	9212	9222
Dec	8621	8363	8480	8279	8289
R ²		.96	.96	.96	.81
RMS		.007	.007	.007	.037
U-coefficient		1.13	1.13	1.14	1.14
$\mathcal{S} = .014$					

Table 3
Observed, Predicted Values of 1969,
and Related Statistics for Regression Equation 3,
in 100,000 of Cubic Feet

	Observed		Predicted		
<u>1969</u>	Y	Y	Y ₁	Y ₂	Y ₃
Jan	1823	1490	1526	1571	1641
Feb	1826	1320	1352	1397	1467
Mar	1546	1267	1297	1342	1412
Apr	1263	1294	1325	1370	1440
May	971	1220	1249	1294	1364
Jun	996	1006	1030	1075	1145
Jul	770	851	871	916	986
Aug	710	905	927	972	1042
Sep	875	1067	1093	1138	1208
Oct	1096	1182	1210	1255	1325
Nov	1571	1273	1304	1349	1419
Dec	1657	1423	1457	1468	1538
R ²		.79	.80	.80	.78
RMS		.027	.027	.027	.027
U-coefficient		.83	.83	.83	.83

$\mathcal{A} = .024$

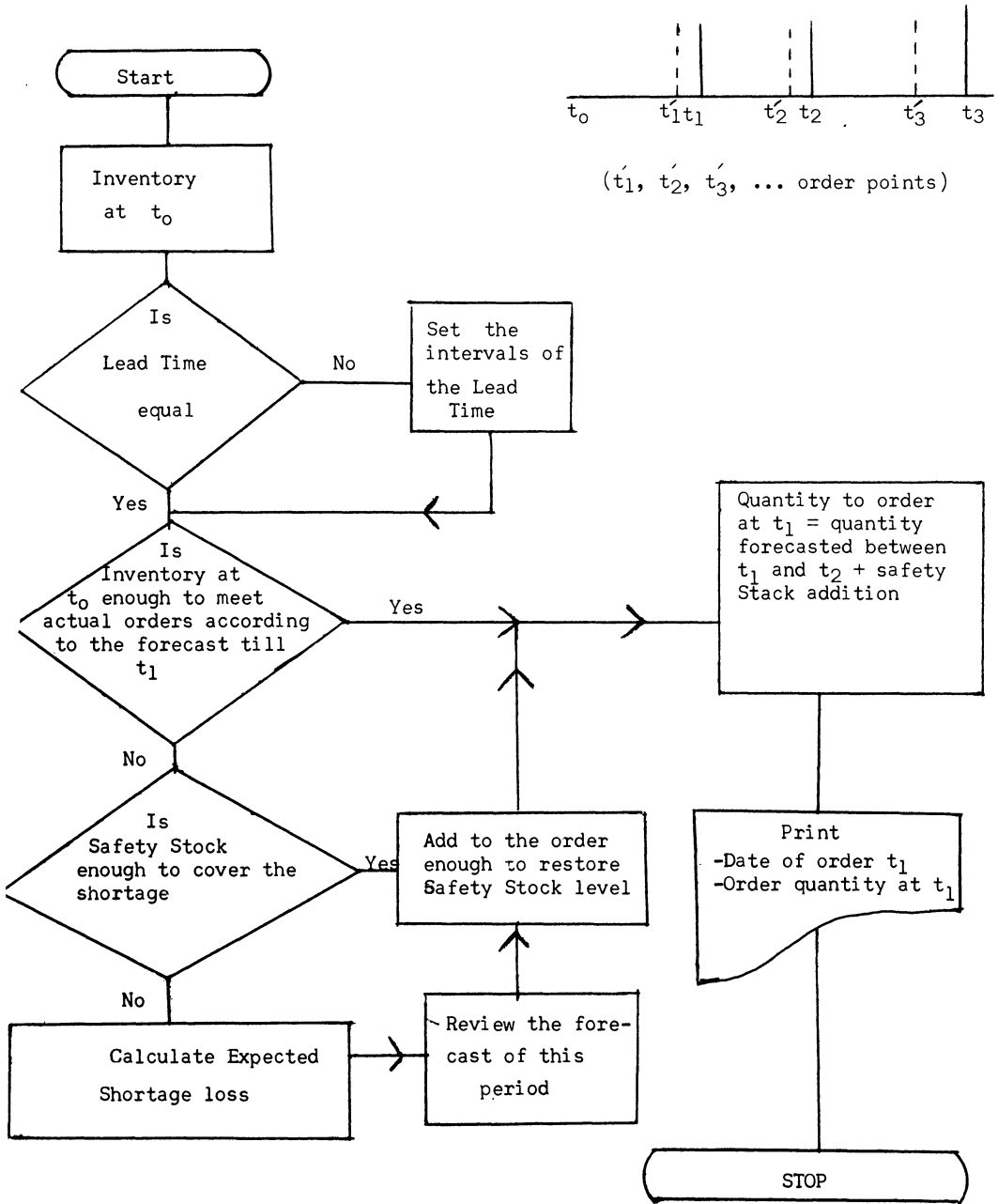
Table 4

Observed, Predicted Values of 1969,
and Related Statistics for Regression Equation 4,
in 100,000 of Cubic Feet

	Observed		Predicted		
<u>1969</u>	Y	Y	Y_1	Y_2	Y_3
Jan	4088	2796	3045	3138	3206
Feb	4096	2623	2856	2949	3017
Mar	3435	2454	2672	2765	2833
Apr	3091	2173	2366	2459	2527
May	1174	1628	1773	1866	1934
Jun	1129	1031	1123	1216	1284
Jul	984	692	754	847	915
Aug	746	648	706	799	867
Sep	805	803	874	967	1128
Oct	1203	1180	1285	1378	1446
Nov	2435	1770	1928	2021	2089
Dec	3683	2391	2604	2697	2765
R^2		.73	.81	.84	.85
RMS		.081	.081	.080	.084
U-coefficient		.68	.68	.68	.70

$\sigma = .089$

Appendix I



Flow chart of an inventory control system (with or without equal lead time)

REFERENCES

1. Allen, R.G.D., Mathematical Economics, London, MacMillan and Co., Ltd., 1966.
2. Bliss, C.I., Periodic Regression in Biology and Climatology, Bulletin 615, New Haven, Connecticut Agricultural Experimental Station, 1958.
3. Eilon, S. and ElMaleh, J., "Adaptive Limits in Inventory Control", Management Science, Vol. 16, No. 8, April 1970.
4. Greene, J.H., Editor-in-Chief, Production and Inventory Control Handbook, New York, McGraw-Hill Book Co., 1970, Chapter 16.
5. Gross, D. and Soriano A., "The Effect of Reducing Leadtime on Inventory Levels - Simulation Analysis", Management Science, Vol. 16, No. 2, Oct. 1969.
6. Hayes, R.H., "Statistical Estimation Problems in Inventory Control", Management Science, Vol. 15, No. 11, July 1969.
7. Packer, A.H., "Simulation and Adaptive Forecasting as Applied to Inventory Control", Operation Research, July 1967.
8. Sirazlian, B.D., "Dimensional and Computational Analysis in Stationary (s,S) Inventory Problems with Gamma Distributed Demand", Management Science, Vol. 17, No. 6, February 1971.

**Spezielle Aspekte
der
mathematischen Programmierung**

Pseudokomplementärverfahren
Zusammenfassung
 von U. Eckhardt, Aachen

Ausgegangen wird von der folgenden Problemstellung:

$$\begin{aligned} &\text{Unter den Nebenbedingungen } \langle a_t, x \rangle \geq b_t \\ &\text{für alle } t \in T \text{ ist } \langle c, x \rangle \text{ zu maximieren.} \end{aligned} \quad (P)$$

Dabei ist $\langle \cdot, \cdot \rangle$ das Skalarprodukt im \mathbb{R}^d , $c \in \mathbb{R}^d$, $a_t \in \mathbb{R}^d$, $b_t \in \mathbb{R}$ für alle $t \in T$, T eine (endliche oder unendliche) Indexmenge.

Das duale Problem im Sinne der "Semi-Infinite Programming Theory" von Charnes, Cooper und Kortanek [1] ist:

$$\begin{aligned} &\text{Gesucht ist eine endliche Teilmenge } I \text{ von } T, \\ &\text{so daß } \sum_{t \in I} y_t \cdot b_t \text{ minimal ist, } \sum_{t \in I} y_t \cdot a_t = c, \\ &y_t = 0 \text{ für } t \in T \setminus I \text{ und } y_t \geq 0 \text{ für alle } t \in I. \end{aligned}$$

Wir nehmen an, es sei eine dual zulässige Basislösung und eine "näherungsweise" zulässige Lösung von (P) bekannt (zur Definition vgl. [4]). Dann kann man ein Verfahren für die Lösung des Problemspaars (P), (D) konstruieren, das auf folgenden Prinzipien beruht:

1. Auf das Problem (D) wird die gewöhnliche Simplex-Methode mit einer speziellen Auswahlvorschrift angewandt.

2. Das Problem (P) wird mit einem Iterationsverfahren gelöst.
3. Die Pivotwahl für das Simplex-Verfahren wird durch die Iterierten des Iterationsverfahrens gesteuert, die jeweilige Fortschreitungsrichtung beim Iterationsverfahren ergibt sich aus den sukzessiven Basislösungen des Simplex-Verfahrens.

Der zusätzliche Rechenaufwand pro Simplex-Schritt ist außerordentlich gering. Es ergeben sich folgende Vorteile:

1. Das Iterationsverfahren benutzt zur Berechnung der Schrittlänge nur die Daten des Ausgangsproblems. Damit ergibt sich eine große numerische Stabilität.
2. Numerische Experimente deuten darauf hin, daß unter gewissen Voraussetzungen beträchtliche Einsparungen an Rechenzeit möglich sind.
3. Bereits vorhandene gute Näherungslösungen können mit Vorteil verwendet werden.

Man kann zeigen, daß das Verfahren in einem Spezialfall mit dem fastkomplementären "Scheme I" von Lemke [6] zusammenfällt.

Die Eigenschaften des Verfahrens legen die Anwendung auf Probleme der numerischen Mathematik nahe, wo man im allgemeinen sehr gute Informationen über das Problempaar (P), (D) hat, insbesondere gute Näherungslösungen oder sogar zulässige Lösungen von (P). Eine Zusammenstellung von Anwendungen der linearen Optimierung in der numerischen Mathematik findet man bei Rabinowitz [7].

Eine Möglichkeit, gute Ausgangsnäherungen für eine zulässige Lösung von (P) zu finden, bieten iterative Verfahren zur Lösung von linearen Ungleichungssystemen [5]. Einzelheiten über das hier skizzierte Verfahren sowie weitere Eigenschaften findet man in [2], [3] und [4].

Literatur:

1. CHARNES, A., W. W. COOPER und K. O. KORTANEK: On the theory of semi-infinite programming and a generalization of the Kuhn-Tucker saddle point theorem for arbitrary convex functions. Naval Res. Logist. Quart. 16, 41 - 51 (1969)
2. ECKHARDT, U.: Eine Modifikation der Auswahlvorschrift beim Simplex-Verfahren. In: R. Henn, H. P. Künzi und H. Schubert, eds.: Operations Research Verfahren VI. Meisenheim: Verlag Anton Hain 1969, S. 93 - 97.
3. ECKHARDT, U.: Fastkomplementäre Iterationspfade und Teilprobleme beim linearen Programmieren. In: R. Henn, H. P. Künzi und H. Schubert, eds.: Operations Research Verfahren VIII. Meisenheim: Verlag Anton Hain 1970, S. 64 - 76.
4. ECKHARDT, U.: Pseudo-complementary algorithms for mathematical programming. In: F. Lootsma, ed.: Numerical Methods for Nonlinear Optimization. London und New York: Academic Press, Inc. 1972, S. 301 - 312.
5. ECKHARDT, U.: Iterative Lösung linearer Ungleichungssysteme. Erscheint als Bericht der KFA Jülich.
6. LEMKE, C. E.: Complementary pivot theory. In: G. B. Dantzig und A. F. Veinott, eds.: Mathematics of the Decision Sciences, Part I. Lectures in Applied Mathematics, Vol. 11. Providence: AMS 1968, S. 95 - 114.
7. RABINOWITZ, P.: Applications of linear programming to numerical analysis. SIAM Rev. 10, 121 - 159 (1968).

Ein Algorithmus zur Konstruktion von Stromzirkulationen in Netzen

von H. Steckhan, Heidelberg

Überblick

Es wird ein neuer Markierungsalgorithmus gezeigt, mit dem man in einem zweiseitig beschränkten Netz eine zulässige Stromzirkulation konstruieren oder nachweisen kann, dass keine solche Zirkulation existiert. Durch dieses Verfahren wird die Suche nach einer zulässigen Basislösung mit Methoden der linearen Programmierung für Netzstromprobleme umgangen.

Wie der alternative Zirkulationsalgorithmus von Ford und Fulkerson kann auch das neue Verfahren leicht in den Out-of-Kilter-Algorithmus integriert werden. Die Anzahl der Iterationen ist im neuen Algorithmus oft kleiner und niemals grösser als im Verfahren von Ford und Fulkerson.

Vorbemerkungen

Viele wohlbekannte Standardprobleme des Operations Research wie

die Konstruktion eines kürzesten oder schnellsten Weges
zwischen zwei Knoten eines (Verkehrs-) Netzes,

die Konstruktion eines maximalen Stromes zwischen Quelle und
Mündung eines beschränkten (Informations- oder Transport-)
Netzes,

das Zuordnungsproblem (Assignmentproblem),

das Transportproblem vom Typ Hitchcock - Kantorowitsch -
Koopmans,

das Transshipmentproblem,

das Tanker - Routing - Problem,

das Catererproblem,

das Lagerhausproblem und andere mehr

können auf die Konstruktion von kostenminimalen zulässigen Zirkulationen in zweiseitig beschränkten Netzen rückgeführt und deshalb einheitlich mit dem Out-of-Kilter-Algorithmus gelöst werden. Dieser Algorithmus, der nach wie vor seine zentrale Position innerhalb der konstruktiven Verfahren der Netzstromanalyse behauptet, kann in seiner einphasigen, extensiven Originalform ([2] pp 162-169, [3]) oder als Zwei-Phasen-Algorithmus eingesetzt werden (Vgl. [2] pp 164-165 mit [2] pp 52-53).

In der Zwei-Phasen-Variante klärt man zunächst mit dem Zirkulationsalgorithmus von Ford und Fulkerson ([2] pp 52-53), ob im gegebenen Netz überhaupt eine zulässige Zirkulation existiert. Fällt dieser Test negativ aus, so erübrigt es sich natürlich, in der leeren Menge der zulässigen Zirkulationen nach einer kostenminimalen zulässigen Zirkulation zu suchen. Erhält man hingegen als einziges alternatives Resultat der ersten Phase eine zulässige Zirkulation, so lässt sie sich als zulässige Ausgangslösung in den Out-of-Kilter-Algorithmus einsetzen (vgl. [2] p 164). In diesem Fall vereinfacht sich der Out-of-Kilter-Algorithmus während der zweiten Phase in drastischer Weise ¹⁾.

Im folgenden zeigen wir eine Alternative zum Zirkulationsalgorithmus von Ford und Fulkerson, entwickeln also eine alternative erste Phase für den zweiphasigen Out-of-Kilter-Algorithmus.

1) Vier der sechs Out-of-Kilter-Zustände (nämlich $\alpha_1, \beta_1, \beta_2$ und γ_2) fallen weg, weil die Anfangszirkulation zulässig ist (vgl. [2] p. 163). Das reduziert die Vielfalt der Markierungsregeln ([2] pp 164-165) in beträchtlichem Ausmass.

Definitionen

Bezeichne (N, A, b, c) ein zweiseitig beschränktes Netz, so dass

$N := \{1, 2, \dots, n\}$, $n > 1$, die Menge der Knoten $x \in N$ symbolisiert

$A := \{(x, y) \in N \times N \mid x \neq y\}$ die Menge der Bögen $(x, y) \in A$,

b die auf die Menge A definierte ganzzahlige Funktion der Mindestströme $b(x, y)$ und

c die auf der Menge A definierte ganzzahlige Funktion der Bogenkapazitäten $c(x, y)$,

wobei $0 \leq b(x, y) \leq c(x, y)$ für alle $(x, y) \in A$.

Die Funktionswerte $f(x, y)$ einer auf der Menge A definierten nicht-negativen Funktion f heissen "Bogenströme" (Strom, der pro Zeiteinheit im Bogen (x, y) vom Knoten x zum Knoten y fliesst).

$f(x, y)$ heisst "zulässiger Bogenstrom", falls $b(x, y) \leq f(x, y) \leq c(x, y)$.

Seien X, Y Teilmengen aus N , sei g eine auf der Menge N definierte Funktion und sei h eine auf der Menge A definierte Funktion, so schreiben wir vereinfachend

$g(X)$ anstelle von $\sum_{x \in X} g(x)$

(X/Y) anstelle von $(X \times Y) \cap A$

$h(X/Y)$ anstelle von $\sum_{(x, y) \in (X \times Y) \cap A} h(x, y)$

und ersetzen in diesen Ausdrücken X durch x und Y durch y , falls X bzw. Y genau ein Element enthalten.

f heisst "Zirkulation, falls $f(x/N) = f(N/x)$ für alle $x \in N$.

Das heisst: "... falls pro Zeiteinheit in jedem Knoten x des Netzes ebenso viel abströmt wie zuströmt."

Eine Zirkulation heisst "zulässig", falls ein jeder ihrer Bogenströme $f(x, y)$ zulässig ist.

Falls X eine echte, nichtleere Teilmenge aus der Menge N der

Knoten $x \in N$ ist und $\bar{X} = N - X$, dann heisst die Teilmenge (X/\bar{X}) aus der Menge A "Schnitt".

Problembeschreibung

Das zu lösende Problem heisst

"Für ein gegebenes zweiseitig beschränktes Netz (N, A, b, c) finde man eine zulässige Zirkulation f oder zeige, dass keine existiert."

Das Hoffmansche Zirkulationstheorem

Eine Teillösung dieses Problems bietet das folgende Theorem
([1], [2] p. 51)

Theorem I : In einem zweiseitig beschränkten Netz (N, A, b, c)
(A. J. Hoffman) existiert eine zulässige Zirkulation f genau dann,
wenn

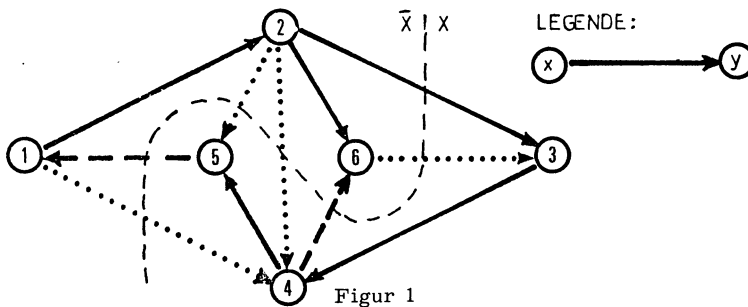
$$c(X/\bar{X}) \geq b(\bar{X}/X)$$

für alle nichtleeren echten Teilmengen X aus N
erfüllt ist.

Das heisst etwa (vgl. die eingeblendete Figur 1):

"... , wenn über den Schnitt (X/\bar{X}) - unterbrochene Bögen - in Richtung X pro Zeiteinheit ebenso viel strömen k a n n, wie umgekehrt über den Schnitt (\bar{X}/X) - gepunktete Bögen - in Richtung X pro Zeiteinheit strömen m u s s."

Das heisst in anderen Worten: "... wenn im 'Bereich' X ein Stromstau und im 'Bereich' \bar{X} ein Sog vermieden werden können." ²⁾



Figur 1

2) Beweise für Theorem I findet man in [1] und [2] pp 50-51.

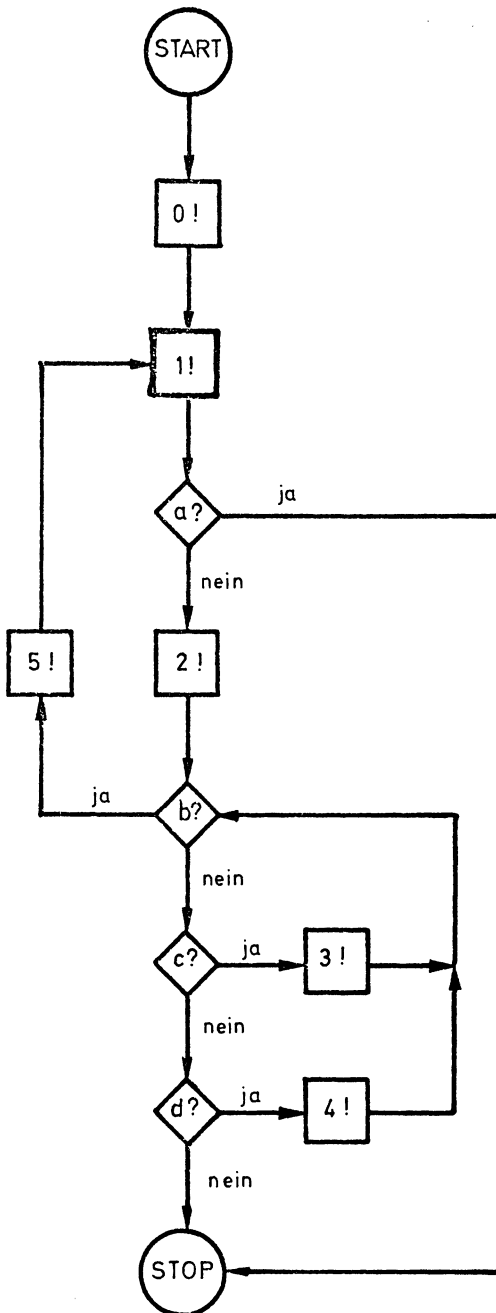
Diesem Theorem zufolge wäre die stattdliche Anzahl von $2^n - 2$ Schnitten zu testen, nur um die Frage der blossen Existenz von zulässigen Zirkulationen zu klären. Überdies bietet das Theorem lediglich eine Teillösung unseres Problems; denn es enthält keine konstruktiven Vorschläge, anhand derer eine zulässige Zirkulation im Falle ihrer Existenz gefunden werden könnte. Deshalb werden wir das Hoffmansche Zirkulationstheorem in späteren Beweisen auch lediglich als notwendige Existenzbedingung verwenden.

Der neue Algorithmus

Neben dem eingeblendeten Flussdiagramm (Figur 2), das den Algorithmus im Überblick zeigt, kann man die Bedeutung der einzelnen Fragen und Anweisungen ablesen (vgl. unten). Wir werden sie sukzessiv anhand eines Beispiels (Figur 3) durcharbeiten, das nun links gezeigt wird. In diesem Beispiel soll jede Frage und jede Anweisung des Flussdiagramms und auch jeder Pfeil mit Ausnahme von "d? \rightarrow STOP" wenigstens einmal durchlaufen werden. Als Resultat entsteht eine kostenminimale zulässige Zirkulation. Würde man hingegen - wie in einem späteren zweiten Beispiel - die Frage "d?" verneinen müssen, so würde keine zulässige Zirkulation existieren.

Im Flussdiagramm bedeuten

- 0! Wähle ganzzahlige, zulässige Bogenströme $f(x, y)$.
- 1! Berechne die "Knotendefizite" (Nettoabströme)
 - $d_f: = f(x/N) - f(N/x)$ für alle $x \in N$.
 - und die Menge $U_f: = \{ x \in N \mid d_f(x) < 0 \}$ der Überschussknoten
- a? Ist U_f leer?
- 2! Markiere einen beliebigen Überschussknoten $p \in U_f$ mit irgend einem Zeichen.
- b? Ist ein Defizitknoten $x \in D_f$ markiert? (Die Frage ist an dieser Stelle sinnvoll, weil sie auch im Anschluss an die Anweisungen 3! und 4! gestellt wird)
- 3! und 4! gestellt wird)
- c? Existiert ein Bogen $(x, y) \in A$, so dass x markiert ist, y unmarkiert und $c(x, y) > f(x, y)$?



Figur 2

- 3! Wähle einen Bogen mit den in "c?" erfragten Eigenschaften aus, nenne ihn (\hat{x}, \hat{y}) und markiere (\hat{x}, \hat{y}) mit "+" und \hat{y} mit " \hat{x} ".
- d? Existiert ein Bogen $(x, y) \in A$, so dass x unmarkiert, y markiert und $f(x, y) > b(x, y)$?
- 4! Wähle einen Bogen mit den in "d?" erfragten Eigenschaften aus, nenne ihn (\check{x}, \check{y}) und markiere (\check{x}, \check{y}) mit "-" und \check{x} mit " \check{y} ".
- 5! Sei q der markierte Defizitknoten. Finde rückwärts, von q aus, anhand der Knotenmarken den Pfad $P_{p, q}$, der von p nach q führt, und diejenigen Bögen als Elemente enthält, auf denen man durch Markierung von p nach q gelangt ist.

Sei V die Menge der von p aus gesehenen Vorwärtsbögen, die auf dem Pfad $P_{p, q}$ liegen und in Richtung q weisen:

$$V = \{ (x, y) \in P_{p, q} \mid (x, y) \text{ ist mit "+" markiert} \}.$$

Sei R die Menge der von p aus gesehenen Rückwärtsbögen, die auf dem Pfad $P_{p, q}$ liegen und in Richtung q weisen:

$$R = \{ (x, y) \in P_{p, q} \mid (x, y) \text{ ist mit "-" markiert} \}.$$

Berechne

$$\begin{aligned} \epsilon_V &= \min_{(x, y) \in V} [c(x, y) - f(x, y)] \\ \epsilon_V &= \infty \text{ wenn } V = \emptyset \\ \epsilon_R &= \min_{(x, y) \in R} [f(x, y) - b(x, y)] \\ \epsilon_R &= \infty \text{ wenn } R = \emptyset \end{aligned}$$

$$\epsilon := \min [\epsilon_V, \epsilon_R, |d_f(p)|, d_f(q)]$$

Bilde neue Bogenströme

$$f'(x, y) = \begin{cases} f(x, y) + \epsilon & \text{wenn } (x, y) \in V \\ f(x, y) - \epsilon & \text{wenn } (x, y) \in R \\ f(x, y) & \text{sonst} \end{cases}$$

Lösche alle Markierungen.

Knoten 4 mit "7". b? wird nun erstmals bejaht. In Anweisung 5! verwenden wir $q=4$.

Die Knotenmarken dienen nun als "Faden der Ariadne." Sie markieren von q aus rückwärts den Pfad $P_{p,q}$, auf dem wir von p nach q gelangt sind:

Da $q=4$ die Marke "7" trägt und nicht der Bogen $(7, 4)$, sondern der Bogen $(4, 7)$ markiert ist, erhalten wir $(4, 7) \in P_{p,q}$. Weil ferner Knoten 7 mit "2", Bogen $(7, 2)$ nicht aber Bogen $(2, 7)$ markiert ist, gilt $(7, 2) \in P_{p,q}$. Schliesslich erhält man $(1, 2) \in P_{p,q}$, weil Knoten 2 die Marke "1" trägt und Bogen $(1, 2)$ markiert ist. Damit gelten $P_{p,q} = \{(1, 2), (7, 2), (4, 7)\}$, $V = \{(1, 2)\}$, $R = \{(7, 2), (4, 7)\}$; $P_{p,q}$ ist also nur eine Teilmenge der markierten Bögen.

Man erhält weiterhin $\epsilon_V = 1$, $\epsilon_R = 2$, $\epsilon = d_f(q) = \epsilon_V = 1$.

Das führt zu den neuen Bogenströmen

$$f'(x, y) = \begin{cases} 8 & \text{wenn } (x, y) = (1, 2) \\ 2 & \text{wenn } (x, y) = (7, 2) \text{ oder } (4, 7) \\ f(x, y) & \text{sonst.} \end{cases}$$

Alle Markierungen werden gelöscht.

Im zweiten und letzten Durchlauf berechnet man im Einklang mit Anweisung 1! die Knotendefizite $d_{f'}(1) = -2$, $d_{f'}(6) = 2$, $d_{f'}(x) = 0$ für alle $x \neq 1, 2$ und $U_{f'} = \{1\}$. a? wird verneint und Knoten $x=1$ mit "§" markiert. Dann werden b? und c? verneint, d? aber bejaht, so dass der Anweisung 3! zufolge Knoten 6 mit "1" und Bogen $(6, 1)$ mit "-" zu markieren sind. b? wird nun bejaht, und man erhält nach Anweisung 5!

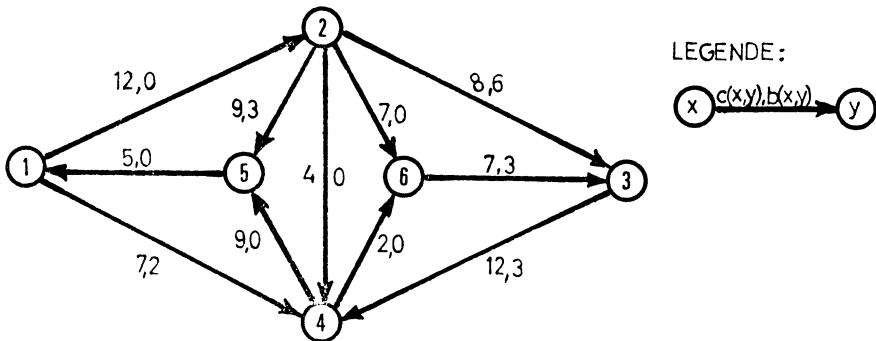
$P_{p,q} = \{(6, 1)\}$, $V = \emptyset$, $R = \{(6, 1)\}$, $\epsilon_V = \infty$, $\epsilon_R = 7$ und $\epsilon = 2$. Das führt zu

$$f''(x, y) = \begin{cases} 7 & \text{wenn } (x, y) = (6, 1) \\ f'(x, y) & \text{sonst.} \end{cases}$$

Im Einklang mit Anweisung 1! ermitteln wir nun $d_{f''}(x) = 0$ für alle $x \in \mathbb{N}$, so dass $U_{f''} = \emptyset$. Folglich wird die Frage a? bejaht.

Wie wir allgemein beweisen werden, bilden die zuletzt definierten

Zweites Beispiel (Nachweis der Nichtexistenz einer zulässigen Zirkulation)



Figur 4

Im zweiten und letzten Beispiel (Figur 4) wählen wir nach Anweisung 0! die zulässigen Bogenströme $f(1, 2) = 3$, $f(1, 4) = 2$, $f(2, 3) = 6$, $f(2, 4) = 0$, $f(2, 5) = 3$, $f(2, 6) = 1$, $f(3, 4) = 9$, $f(4, 5) = 2$, $f(4, 6) = 2$, $f(5, 1) = 5$ und $f(6, 3) = 3$. Daraus ergeben sich nach Anweisung 1! $d_f(2) = 7$, $d_f(4) = -7$, $d_f(x) = 0$ wenn $x \neq 2, 4$ und $U_f = \{4\}$. Frage a? wird verneint, so dass der Knoten 4 mit "§" markiert wird. Nachdem b? verneint und c? bejaht worden sind, markieren wir in Übereinstimmung mit Anweisung 3! den Bogen (4, 5) mit "+" und den Knoten 5 mit "4". Nun werden b? und c? verneint, d? hingegen bejaht. Der Anweisung 4! zufolge sind der Bogen (3, 4) mit "-" und der Knoten 3 mit "4" zu markieren. Schliesslich werden die Fragen b?, c? und d? der Reihe nach verneint, und der Algorithmus endet mit dem beweispflichtigen Resultat, dass im gegebenen zweiseitig beschränkten Netz (N, A, b, c) deshalb keine zulässige Zirkulation existiert, weil die Hoffmansche Existenzbedingung für die Menge X der markierten Knoten unerfüllt bleibt (vgl. hierzu Figur 1).

Beweise zum neuen Algorithmus

Satz 1 Der neue Algorithmus endet nach endlich vielen endlichen Schritten (Stromänderungen) entweder mit der Verneinung der Frage $d?$ oder mit der Bejahung der Frage $a?$

Beweis

Da ein zweiseitig beschränktes Netz (N, A, b, c) nur endlich viele nämlich n Knoten enthält, und mit jedem Durchlaufen der Schleifen $b? - c? - 3! - b?$ und $b? - c? - d? - 4! - b?$ ein Knoten markiert wird, gelangt man spätestens nach n Knotenmarkierungen entweder zur Verneinung der Frage $d?$ (dann können keine weiteren Knoten markiert werden) oder zur Bejahung der Frage $b?$. Wird aber $b?$ bejaht, so folgt eine Stromänderung. Folglich ist jeder Schritt endlich.

ϵ ist nach Konstruktion eine strikt positive ganze Zahl. Folglich vermindert sich $|d_f(U_f)|$, das als Summe von strikt positiven ganzen Zahlen $d_f(x)$ selbst eine strikt positive ganze Zahl ist, mit jeder Stromänderung um wenigstens 1. Deshalb führen nicht mehr als $|d_f(U_f)|$, also endlich viele, Stromänderungen entweder zur Bejahung der Frage $a?$ oder zur Verneinung der Frage $d?$

Satz 2 Wenn die Frage $a?$ bejaht wird, dann bilden die zuletzt definierten Bogenströme eine zulässige Zirkulation.

Beweis

Da die anfangs gewählten Bogenströme zulässig sind, beweist man Satz 2, indem man zeigt, dass

- (1) jede Stromänderung zulässige Bogenströme ergibt und
- (2) die Zirkulationseigenschaft

$$f(x/N) = f(N/x) \text{ für alle } x \in N$$

erfüllt wird, wenn $U_f = \emptyset$.

(1) folgt unmittelbar aus der Konstruktion des ϵ und der $f'(x, y)$.

Sei $G_f = \{ x \in N \mid d_f = 0 \}$ die Menge der ausgeglichenen Knoten, und sei $D_f = \{ x \in N \mid d_f > 0 \}$ die Menge der Defizitknoten, dann sind die

Mengen U_f , D_f , G_f paarweise elementefremd und es gilt

$U_f \cup D_f \cup G_f = N$. Daraus folgt

$$(3) \quad f(N/N) = f(D_f/N) + f(U_f/N) + f(G_f/N) = f(N/D_f) + f(N/U_f) + f(N/G_f).$$

Die Definition für $d_f(x)$ ergibt

$$f(x/N) = f(N/x) \text{ für alle } x \in G_f.$$

Das führt mit Summierung über alle $x \in G_f$ zu

$$f(G_f/N) = f(N/G_f).$$

daraus folgt mit (3)

$$f(D_f/N) - f(N/D_f) = f(N/U_f) - f(U_f/N).$$

das führt zu

$$(4) \quad d_f(D_f) = |d_f(U_f)|.$$

daraus ergibt sich mit

$$d_f(x) = \begin{cases} > 0 & \text{wenn } x \in D_f \\ < 0 & \text{wenn } x \in U_f, \end{cases}$$

dass D_f genau dann leer ist, wenn U_f leer ist. Da nun N in die Mengen U_f , D_f , G_f zerlegt ist, folgt weiter, dass $G_f = N$ dann und nur dann gilt, wenn $U_f = \emptyset$.

Das führt mit der Definition für G_f dazu, dass $f(x/N) = f(N/x)$ für alle $x \in N$ erfüllt wird, wenn $U_f = \emptyset$.

Folglich gilt auch (2).

Satz 3 Wenn die Frage $d?$ verneint wird, dann existiert keine zulässige Zirkulation.

Beweis

Sei X die Menge der markierten Knoten, wenn $d?$ verneint wird ($X \neq \emptyset$ gilt, weil p markiert ist). wird $d?$ verneint, dann hat man

$$f(x, y) = \begin{cases} c(x, y) & \text{wenn } (x, y) \in (X/\bar{X}) \\ b(x, y) & \text{wenn } (x, y) \in (\bar{X}/X); \end{cases}$$

denn anderenfalls könnten weitere Bögen nach einer der Anweisungen 3! oder 4! markiert werden.

Das führt zu

$$(5) \quad f(X/\bar{X}) = c(X/\bar{X}) \text{ und } f(\bar{X}/X) = b(\bar{X}/X).$$

Wird $d?$ verneint, so gilt nach Konstruktion $D_f \cap K = \emptyset$, weil kein Defizitknoten markiert werden konnte. Daraus folgt:

(6) $f(N/x) \geq f(x/N)$ für alle $x \in X$.

Wird d? verneint, so ist $p \in U_f$ markiert. Folglich hat man

(7) $f(N/x) > f(x/N)$ für $x=p$, $p \in X$.

Aus (6) und (7) folgt durch Summierung über alle $x \in X$, dass

(8) $f(N/X) > f(X/N)$.

Das führt mit $X \cup \bar{X} = N$ und $X \cap \bar{X} = \emptyset$ zu

(9) $f(X/X) + f(\bar{X}/X) > f(X/X) + f(X/\bar{X})$,

so dass

(10) $f(\bar{X}/X) > f(X/\bar{X})$ wenn d? verneint wird.

(5) und (10) ergeben

$$c(X/\bar{X}) < b(\bar{X}/X),$$

was der Hoffmanschen Bedingung widerspricht; deren Erfüllung für die Existenz einer zulässigen Zirkulation notwendig ist.

Vergleich der maximalen Schrittzahlen im Zirkulationsalgorithmus von Ford und Fulkerson und im neuen Algorithmus

Dieser Vergleich wird dadurch erschwert, dass die maximale Schrittzahl in beiden Algorithmen durch die Wahl der anfänglichen Bogenströme beeinflusst wird. Der Rahmen für diese Wahl scheint indessen recht weit gefasst zu sein:

Ford und Fulkerson schreiben lediglich vor "Start with any integral valued f that satisfies the conservation equations (das heisst: $f(N/x) = f(x/N)$) at all nodes. For example $f = 0$ (das heisst: $f(x, y) = 0$ für alle $(x, y) \in A$) will do". ([2] p 52.)

Wir fordern hingegen zu Anfang nichts anderes als ganzzahlige, zulässige Bogenströme und könnten hinzufügen " $f(x, y) = b(x, y)$ für alle $(x, y) \in A$ reicht beispielsweise aus".

Für das Verfahren von Ford und Fulkerson zeigt sich freilich bei näherer Betrachtung, dass $f = 0$ die einzige Anfangszirkulation ist, die im allgemeinen Fall ohne zusätzliche Hilfsalgorithmen gefunden werden kann. Deshalb legen wir sie unserem Vergleich zugrunde. Für den oben beschriebenen neuen Algorithmus

beschränken wir die Auswahl der anfänglichen Bogenströme auf

$$f(x, y) = b(x, y) \text{ für alle } (x, y) \in A.$$

Um Verwechslungen im folgenden Vergleich zu vermeiden, bezeichnen wir die zu Anfang festgelegten Bogenströme im Ford-Fulkerson - Algorithmus mit $F(x, y)$ und in unserem Algorithmus mit $f(x, y)$. Wir gehen also davon aus, dass

$$\begin{aligned} f(x, y) &= b(x, y) \text{ für alle } (x, y) \in A \text{ und} \\ F(x, y) &= 0 \quad \text{für alle } (x, y) \in A. \end{aligned}$$

Seien $Q = \{ (x, y) \in A \mid c(x, y) < F(x, y) \}$ und
 $S = \{ (x, y) \in A \mid F(x, y) < b(x, y) \}$

hilfsweise für die Behandlung des Ford - Fulkerson - Algorithmus definiert.

Bezeichne ferner

w die maximale Schrittzahl im Ford - Fulkerson Verfahren und
 z die maximale Schrittzahl in unserem neuen Algorithmus.

Aus [2] p 53 ergibt sich allgemein

$$(11) \quad w = \sum_{(x, y) \in Q} (F(x, y) - c(x, y)) + \sum_{(x, y) \in S} (b(x, y) - F(x, y)).$$

Diese obere Schranke wird auch tatsächlich erreicht: Wählt man

$N = \{ 1, 2, 3 \}$, $A = \{ (1, 2), (2, 1), (2, 3), (3, 1) \}$, $b(1, 2) = 5$,

$b(2, 1) = b(2, 3) = 0$, $b(3, 1) = 4$, $c(1, 2) = 7$, $c(2, 1) = 1$,

$c(2, 3) = c(3, 1) = 6$ und $F(x, y) = 0$ für alle $(x, y) \in A$, wählt man

ferner ohne Widerspruch zu den Regeln in [2] pp 52-53 die Bögen

$(2, 1)$, $(3, 1)$ in dieser Reihenfolge alternierend als Bogen (s, t)

und markiert man überdies den Bogen $(2, 1)$ jedesmal als zweiten

Bogen, wenn $(s, t) = (3, 1)$ gesetzt wird, dann ist der Strom im

Bogen $(3, 1)$ erst nach dem Ende des achten Schrittes zulässig,

und man benötigt noch einen neunten für den Bogen $(1, 2) = (s, t)$.

Da anfangs $F(x, y) = 0$ für alle $(x, y) \in A$ gewählt wird, und

$c(x, y) \geq b(x, y) \geq 0$ für alle $(x, y) \in A$ gilt, folgt $Q = \emptyset$.

Das führt mit (11) zu

$f(x, y) = 0$ für alle $(x, y) \in A$ und $b(x, y) \geq 0$ für alle $(x, y) \in A$ ergeben

$$b(x, y) - f(x, y) = 0 \text{ für alle } (x, y) \in (A - S).$$

Daraus folgt mit (12), dass

$$w = \sum_{(x, y) \in A} (b(x, y) - f(x, y)).$$

Das führt mit $f(x, y) = 0$ für alle $(x, y) \in A$ zu

$$(13) \quad w = \sum_{(x, y) \in A} b(x, y) = b(N/N).$$

Gegen Ende des Beweises für Satz 1 wurde gezeigt, dass

$$z = |d_f(U_f)|.$$

Das führt mit (4) zu

$$(14) \quad z = d_f(D_f) = \sum_{x \in D_f} (f(x/N) - f(N/x)) = f(D_f/N) - f(N/D_f).$$

Da nun $f(x, y) = b(x, y)$ für alle $(x, y) \in A$ gewählt wird, führt (14) zu

$$z = b(D_f/N) - b(N/D_f).$$

Daraus folgt mit $b(x, y) \geq 0$ für alle $(x, y) \in A$

$$(15) \quad z \leq b(D_f/N).$$

Gegen Ende des Beweises für Satz 2 wurde gezeigt, dass D_f dann und nur dann leer ist, wenn auch U_f leer ist. Daraus folgt, dass D_f stets eine echte Teilmenge aus N darstellt. Berücksichtigt man dies und $b(x, y) \geq 0$ für alle $(x, y) \in A$ im Vergleich zwischen (13) und (15), so folgt

$$(16) \quad z \leq b(D_f/N) \quad \begin{cases} = b(N/N) = w & \text{wenn } b(x, y) = 0 \text{ für alle } (x, y) \in A \\ < b(N/N) = w & \text{sonst} \end{cases}$$

Wenn aber $b(x, y) = 0$ für alle $(x, y) \in A$, dann ist das hier behandelte Problem ohnehin trivial: In beiden Algorithmen bilden dann die anfangs gewählten Bogenströme offensichtlich eine zulässige Zirkulation.

In allen anderen, nicht trivialen Fällen, in denen also $b(x, y) = 0$ nicht für alle $(x, y) \in A$ erfüllt wird, führt (16) zu

$$z < w,$$

was besagt, dass die maximale Anzahl der Schritte im Verfahren von Ford - Fulkerson in allen nicht trivialen Fällen echt grösser ist als in unserem neuen Verfahren.

Zitierte Literatur

- [1] Hoffman, A. J. : Some Recent Applications of the Theory of
Linear Inequalities to Extremal Combinatorial
Analysis. In: Proceedings of Symposia on
Applied Mathematics, Band 10 (1960)
- [2] Fulkerson, D. R. : An Out-of-Kilter Method for Minimal Cost Flow
Problems. In: Journal of the Society of
Industrial and Applied Mathematics, Band 9
(1961), pp 18-27
- [3] Ford, L. R. und D. R. Fulkerson: Flows in Networks, Princeton 1962

Über die Lösung des dreidimensionalen Transportproblems

von W. Junginger, Stuttgart

Zusammenfassung:

Das dreidimensionale Transportproblem (TP3) ist zunächst eine formale Erweiterung des gewöhnlichen Transportproblems auf ein Problem mit dreifach indizierten Variablen x_{ijk} . Verschiedene Fragestellungen lassen sich in dieser Form bequem darstellen. Bei der Lösung des TP3 kann man - ähnlich wie beim gewöhnlichen Transportproblem - infolge der speziellen Nebenbedingungen Vereinfachungen gegenüber den Verfahren für das allgemeine LP finden. Möglichkeiten hierfür zeigt HALEY, doch sind die wesentlichen Teile seines Vorschlags für ein praktisches Vorgehen unbrauchbar. Dadurch wird **aber** die praktische Bedeutung dieses Vorschlags in Frage gestellt. Der Vortrag zeigt nun, wie man durch Einführung eines geeigneten Graphen die fraglichen Teile bequem lösen kann. Dies gilt allerdings zunächst nur solange, als die Basis Dreiecksform besitzt. Aber auch wenn dies nicht erfüllt ist, läßt sich das Verfahren so modifizieren, daß auch dann die notwendigen Schritte zur Lösung des TP3 anhand dieses Graphen durchführbar sind.

1. Einleitung

Das dreidimensionale Transportproblem ist zunächst eine formale Erweiterung des gewöhnlichen Transportproblems auf ein entsprechendes Problem mit dreifach indizierten Variablen: gefragt ist nach den Werten der Variablen

$$x_{ijk} \geq 0 \quad i \in I, k \in K, j \in J \quad (1)$$

die die Zielfunktion

$$z = \sum_i \sum_j \sum_k c_{ijk} x_{ijk} \quad (2)$$

minimieren. Dabei sind I, J, K die jeweiligen Indexbereiche. Für die Nebenbedingungen sind verschiedene Möglichkeiten denkbar. Diesbezügliche Untersuchungen von BARTH und HALEY haben aber gezeigt, daß nur die folgenden Nebenbedingungen ein wesentlich neues Problem ergeben:

$$\sum_i x_{ijk} = a_{jk} \quad j \in J, k \in K \quad (3a)$$

$$\sum_j x_{ijk} = b_{ki} \quad k \in K, i \in I \quad (3b)$$

$$\sum_k x_{ijk} = e_{ij} \quad i \in I, j \in J \quad (3c)$$

Für die Lösung dieses dreidimensionalen Transportproblems (TP3) kommen zunächst die allgemeinen Methoden der linearen Optimierung in Frage. HALEY zeigte aber, daß auch hier wie beim gewöhnlichen Transportproblem (TP2) infolge der speziellen Struktur der Nebenbedingungen ein einfacheres Lösungsverfahren möglich ist. Durch formales Übertragen der MODI - Methode auf den dreidimensionalen Fall kommt HALEY auf ein Verfahren zur Lösung des TP3, das in Abb.1. als Ablaufschema dargestellt ist. Dieses unterscheidet sich prinzipiell nicht von einem entsprechenden Ablaufschema für das TP2. Jedoch bestehen vereinzelt grundsätzliche Unterschiede zwischen TP2 und TP3. Während beim TP2 die Existenz

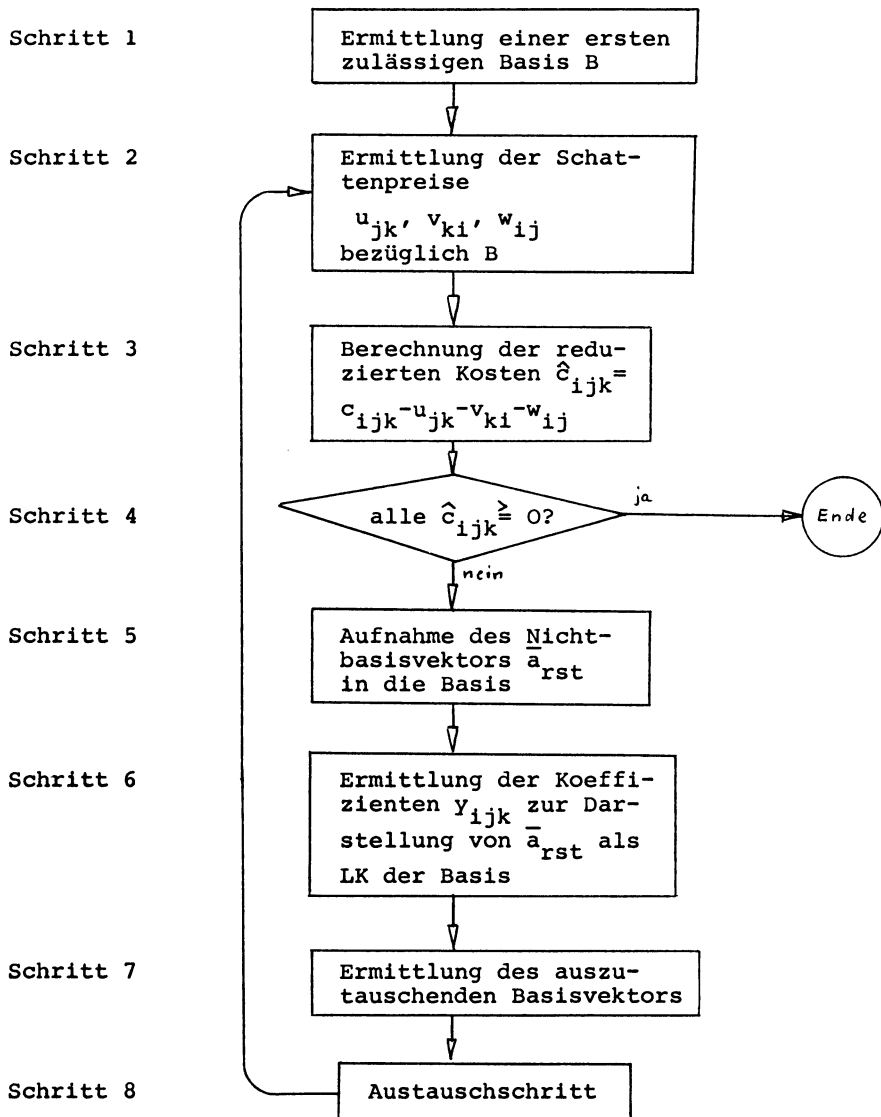


Abb.1: Ablaufschema für das TP3

einer zulässigen Basis immer garantiert ist (Schritt 1), trifft dies auf das TP3 nicht zu. Selbst wenn in (3) die für die a_{jk} , b_{ki} , c_{ij} notwendigen Bedingungen

$$\sum_j a_{jk} = \sum_i b_{ki} \quad k \in K \quad (4a)$$

$$\sum_k b_{ki} = \sum_j e_{ij} \quad i \in I \quad (4b)$$

$$\sum_i e_{ij} = \sum_k a_{jk} \quad j \in J \quad (4c)$$

erfüllt sind, braucht für das TP3 keine zulässige Lösung zu existieren. HALEY gibt eine Möglichkeit an, wie auf jeden Fall entweder eine zulässige Lösung gefunden werden kann oder die Anzeige, daß es eine solche nicht gibt. Deshalb bleibt Schritt 1 im folgenden außer Betracht.

Für die Ermittlung der Schattenpreise in Schritt 2 ist im wesentlichen das Gleichungssystem

$$u_{jk} + v_{ki} + w_{ij} = c_{ijk} \quad \forall ijk \in I_B \quad (5)$$

zuständig; dabei umfaßt I_B alle Indextrippel, die zu den Basisvektoren gehören. Hat man aus (5) die Schattenpreise gewonnen, so bereiten die Schritte 3 bis 5 keine Schwierigkeit. Bei Schritt 6 ist wieder die Auflösung eines Gleichungssystem

$$\bar{B}\bar{y} = -\bar{a}_{rst} \quad (6)$$

notwendig; \bar{B} ist die aus den Basisvektoren gebildete Matrix. Mit der Lösung von (6) sind die Schritte 7 und 8 wieder relativ leicht zu erledigen, man verfährt ganz entsprechend wie beim TP2. Versucht man nun, nach diesem Vorschlag HALEYs ein TP3 durchzurechnen, so kommt man bei den Schritten 2 und 6 in Schwierigkeiten. Denn die Vorschläge, die HALEY zur Lösung der hier benötigten Gleichungssysteme (5) und (6) macht, sind für ein praktisches Vorgehen unbrauchbar. Da aber gerade diese beiden Schritte die wesentliche Arbeit enthalten, war es notwendig, hier neue Möglichkeiten zu finden.

Auch beim TP2 steht man vor der Auflösung von Gleichungssystemen der Art (5) und (6), jedoch haben hier diese Systeme stets Dreiecksgestalt. Dies hat zur Folge, daß bei (5) die Schattenpreise in einem sukzessiven Auflösungsprozeß ermittelt werden können und bei (6) die y_{ij} nur Werte 0, 1 und -1 annehmen; diese Werte werden dann anhand einer geeigneten Schleife im Schema des TP2 gefunden.

Beim TP3 brauchen die Gleichungssysteme (5) und (6) keine Dreiecksform zu haben. Dadurch lassen sich die einfachen Verhältnisse beim TP2 nicht mehr unbedingt übertragen. Aber auch selbst in den Fällen, wo (5) und (6) dreieckig sind, wird eine formale Übertragung dadurch erschwert, daß ein entsprechendes Schema fürs TP3 dreidimensional ist. Nachdem bereits im zweidimensionalen Fall ein effektiver Algorithmus zur Ermittlung der am Austausch beteiligten Basisvariablen (dies entspricht der erwähnten Schleifenbildung für Schritt (6)) nur schwierig zu finden ist, gilt dies in erhöhtem Maße für den dreidimensionalen Fall.

Das Problem besteht also darin, einfache Lösungsmöglichkeiten für die Schritte 2 und 6 zu finden. Ein Ansatz hierfür bietet die Übertragung eines Vorgehens beim TP 2 nach MÜLLER - MERBACH, die im Falle dreieckiger Gleichungssysteme zu guten Ergebnissen führt. Haben jedoch (5) und (6) nicht mehr Dreiecksform, so versagt dieses Vorgehen zunächst. Als weiteres Problem ist deshalb das Lösungsverfahren auch auf diesen Fall zu erweitern.

2. Kalkül zur graphentheoretischen Behandlung des TP3.

Zunächst benötigt man - entsprechend wie beim TP2 - ein geeignetes Schema für das TP3, in dem die zum TP3 gehörigen Daten sowie die Basis dargestellt werden können.

Definition 1 : Unter einem TP3 - Schema versteht man das kartesische Produkt $I \times J \times K$. Die Gesamtheit aller Elemente, die in den Werten jeweils zweier Indizes übereinstimmen, bezeichnet man als Linie des TP3 - Schemas.

Satz 1: Sei \bar{A} die Koeffizientenmatrix des Systems der Nebenbedingungen (3). Ordnet man dann der zur Variable x_{ijk} gehörigen Spalte von \bar{A} das Element (i,j,k) des TP3 - Schemas zu, so ist damit eine eindeutige Zuordnung sowohl der Spalten von \bar{A} zu den

Elementen des TP3 - Schemas wie auch der Zeilen von \bar{A} zu den Linien im TP3 - Schema gegeben.

Wesentlich für die Lösung eines TP3 ist nun, daß man in dem Schema z.B. eine Basis darstellen kann. Man erreicht das durch folgende

Definition 2: Sei $D = \{\bar{a}_{ijk} \mid \bar{a}_{ijk} \in \bar{A}, ijk \in I_D\}$ eine Menge von Spaltenvektoren aus \bar{A} . Dann liegt eine Darstellung von D im TP3 - Schema vor, wenn die Elemente $I_D \subset I \times J \times K$ vor den anderen Elementen ausgezeichnet sind. Ein solches ausgezeichnetes Element wird als Punkt bezeichnet, ihre Gesamtheit mit P_D .

Ein TP3 - Schema kann etwa in der in Abb. 2 angegebenen Weise dargestellt werden. Bei diesem Beispiel sind $I = J = K = \{1, 2, 3\}$. Die Elemente von $I \times J \times K$ werden repräsentiert durch die Positionen längs der Diagonalen in den großen Quadraten. Abb. 2 enthält gleichzeitig die Darstellung einer bestimmten Menge D nach Def. 2; die Punkte sind durch kleine Kreise angegeben.

	$j = 1$	2	3	$k =$
$i = 1$				1 2 3
2				1 2 3
3				1 2 3

Abb. 2: TP3 - Schema mit Darstellung von D

Von besonderer Bedeutung sind nun solche Mengen D , die zu Gleichungssystemen von Dreiecksform führen.

Definition 3: Eine Menge D nach Def. 2 heißt dreieckig, wenn die aus den Vektoren von D gebildete Matrix D durch geeignete Permutationen der Zeilen und Spalten auf die Form einer Dreiecksmatrix mit Diagonalelementen $\neq 0$ gebracht werden kann.

$$\text{Sei} \quad \bar{D} \bar{x} = \bar{b} \quad (7)$$

ein Gleichungssystem, das zu einer dreieckigen Menge D gehört. Dann ist es möglich, dieses Gleichungssystem sukzessive aufzulösen: stets wird es mindestens eine Gleichung geben, in der nur noch eine Variable unbekannt ist. Aus dieser Gleichung kann dann diese Variable ermittelt werden. Ordnet man nun jeweils diese Variable je einer Gleichung zu, so erhält man eine eindeutige Zuordnung z der Variablen von \bar{x} zu einer Teilmenge der Gleichungen von (7). Nach Satz 1 hat diese Zuordnung ihre Entsprechung im TP3-Schema: z induziert eine eindeutige Zuordnung von P_D in die Linien des TP3-Schemas. Diese Zuordnung kann ohne Zuhilfenahme von (7) auch unmittelbar anhand des TP3-Schemas durch sukzessives Zuordnen der Punkte aus P_D zu den Linien im TP3-Schema ermittelt werden. Damit ist jedes $P_i \in P_D$ genau einer Linie $z(P_i)$ des TP3-Schemas zugeordnet.

Aus der Dreieckseigenschaft von D folgt, daß $z(P_D)$ nicht sämtliche Linien des TP3-Schemas umfaßt. Folgende Klassifizierung ist deshalb zweckmäßig:

Definition 4: Sei D dreieckig und z eine zugehörige Zuordnung.

Dann heißen die Linien $z(P_D)$ Linien mit Zuordnung, die übrigen Linien heißen Endlinien.

Mit Hilfe von z wird nun zu D ein Graph definiert, mit dessen Hilfe die Auflösung von (7) möglich ist.

Definition 5: D sei eine dreieckige Menge und P_D die zur Darstellung von D im TP3-Schema gehörigen Punkte; ferner sei z eine zu D gehörige Zuordnung. Dann ist der zu z gehörende Graph von D der gerichtete Graph $G = \{P_D, E \subset P_D \times P_D\}$, wobei genau dann $P_1 P_2 \in E$ eine Kante ist, wenn P_1 und P_2 auf derselben Linie liegen und diese P_2 zugeordnet ist.

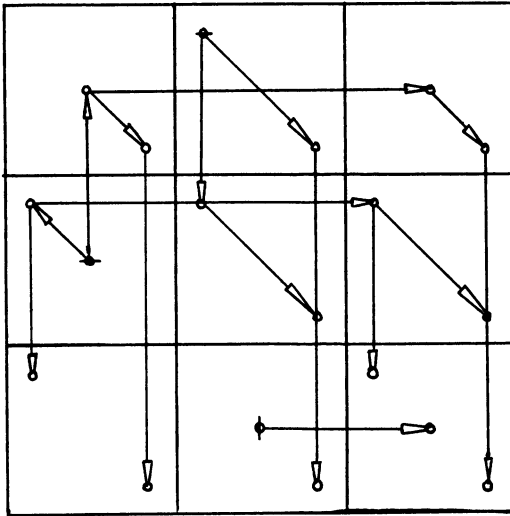


Abb. 3: Der zu z gehörende Graph von D

Abb.3 zeigt einen zu der in Abb.2 angegebenen Menge D möglichen Graphen G. Er wird als Erweiterung der Darstellung von D im TP3-Schema gezeichnet, indem man noch die Kanten mit hinzunimmt. Kanten gehen dabei stets von einem Punkt P_i zu solchen Punkten P_j , die einer der drei Linien durch P_i zugeordnet sind. Münden dabei mehrere Kanten aus derselben Richtung in einen Punkt ein, so wird in Abb.3 nur die jeweils umfassendere eingezeichnet.

Für eine Lösung von (7) anhand G ist wesentlich, daß G keine Zyklen enthält.

Satz 2: Sei G der zu z gehörende Graph von D nach Def.5. Wurde dann z in der oben angegebenen Weise durch sukzessives Zuordnen ermittelt, so enthält G keine Zyklen.

Wesentlich ist hier, daß das sukzessive Zuordnen entsprechend einer sukzessiven Auflösung von (7) geschieht. In diesem Fall heißt z selbst auch zyklensfrei. Grundsätzlich kann auch eine Zuordnung z von P_0 in die Linien des TP3 - Schemas ohne sukzessiven Prozeß ermittelt werden. In diesem Fall kann aber G Zyklen enthalten; z braucht dann nicht mehr zyklensfrei zu sein. Sei G zyklensfrei. Dann ist folgende Definition sinnvoll;

Definition 6: $G = \{P_D, E\}$ sei endlich und **zyklenfrei**. Gibt es dann zu $P_1 \in P_D$ ein $P_2 \in P_D$ mit $P_1 P_2 \in E$, so heißt P_2 Nachfolger 1. Ordnung von P_1 . Ist allgemein $P_j \in P_D$ Nachfolger j -ter Ordnung von P_1 und P_k Nachfolger 1. Ordnung von P_j , so heißt P_k Nachfolger $j+1$ -ter Ordnung von P_1 . P_1 selbst heißt Nachfolger 0-ter Ordnung von sich selbst. Die Gesamtheit aller Nachfolger von P_1 heißt Nachfolgerbereich von P_1 .

Definition 7: Sei G endlich und zyklenfrei. Gibt es dann in G zu P_0 ein P_i , für das P_0 Nachfolger i -ter Ordnung ist, so heißt P_i Vorgänger i -ter Ordnung von P_0 . Die Gesamtheit aller Vorgänger von P_0 heißt Vorgängerbereich von P_0 .

3. Graphentheoretische Lösung spezieller Gleichungssysteme.

Mit Hilfe des in 2. entwickelten Kalküls kann die Lösung der Gleichungssysteme (5) und (6) in Angriff genommen werden. Hierzu sei D eine dreieckige Menge nach Def.2 und \bar{D} die zugehörige Matrix der Spaltenvektoren. Dann hat das Gleichungssystem

$$\bar{D} \bar{y} = -\bar{a}_{rst} \quad (8)$$

dieselbe Form wie (6). Zu seiner Lösung wird noch eine zyklensfreie Zuordnung z von D sowie der zugehörige Graph G benötigt. G ist dann zyklensfrei.

Definition 8: Sei D eine dreieckige Menge, z eine Zyklensfreie Zuordnung von D und G der zugehörige Graph. Unter der Einschränkung von (8) auf G versteht man dann die Gesamtheit derjenigen Gleichungen von (8), die entsprechend Satz 1 zu den Linien mit Zuordnung von G gehören. Sie wird bezeichnet mit

$$\bar{D}^E \bar{y}^E = -\bar{a}_{rst}^E \quad (9)$$

\bar{D}^E ist eine quadratische Matrix und als Teilmatrix von \bar{D} selbst einer Dreiecksmatrix äquivalent. Deshalb ist (9) sicher lösbar.

Satz 3: Sei z eine zyklensfreie Zuordnung der dreieckigen Menge D und $G = \{P_D, E\}$ der zugehörige Graph. Dann erhält man eine Lösung von (9) in folgender Weise: Zuerst wird den Punk-

ten, die den durch (r, s, t) gehenden Linien mit Zuordnung zugewiesen sind, der Wert -1 zugewiesen, und allen anderen Punkten der Wert 0 . Dann ermittelt man für jeden Punkt mit Wert -1 die Nachfolger der Ordnung j ($j = 1, 2, \dots$) und addiert zum Wert eines solchen Nachfolgers den Wert $(-1)^{j+1}$. Setzt man dann $x_{ijk} = f(P_k)$, wo P_k der zu $\bar{a}_{ijk} \in D$ eindeutig bestimmte Punkt in P_D bedeutet und $f(P_k)$ der bei diesem Prozeß dem Punkt P_k zugewiesene Wert, so bildet die Gesamtheit dieser Werte eine Lösung \bar{y}^E von (9).

Die so erhaltene Lösung ist zunächst nur Lösung von (9). Ist aber (8) lösbar, d.h. gilt $R(\bar{D}) = R(\bar{D}, \bar{b})$, so ist die gefundene Lösung gleichzeitig auch Lösung von (8).

Abb. 4 zeigt als Beispiel für die Anwendung dieses Satzes den Graphen aus Abb. 3, anhand dem nach Satz 3 die Lösung von

$$\bar{D} \bar{y} = -\bar{a}_{222}$$

ermittelt wurde (die Lösung ist hier nicht nur Lösung der Einschränkung, da $R(\bar{D}) = R(\bar{D}, -\bar{a}_{222})$ gilt). Zu den durch $(2, 2, 2)$ gehenden Linien mit Zuordnung gehören hier die Punkte P_1 , P_2 und P_3 . Die Wertzuweisungen 1 und -1 sind bei den einzelnen Punkten mit $+$ und $-$ markiert; die Summe dieser Markierungen ergeben dann die gesuchten Werte der Lösung $\bar{y}^E = \bar{y}$.

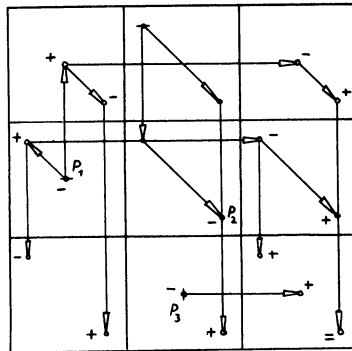


Abb. 4: Lösung des Gleichungssystems $\bar{D} \bar{y} = -\bar{a}_{222}$

Als nächstes wird das Gleichungssystem

$$\bar{D}' \bar{y} = \bar{c} \quad (10)$$

betrachtet, wo \bar{D}' die Transponierte von \bar{D} ist und \bar{c} irgend ein Vektor bedeutet. Bei geeigneter ⁶Setzung von \bar{y} durch einen Vektor mit Komponenten u_{jk}, v_{ki}, w_{ij} wird (10) zu

$$u_{jk} + v_{ki} + w_{ij} = c_{ijk} \quad \forall ijk \in I_D \quad (11)$$

und hat damit dieselbe Form wie (5).

Die Lösung von (11) geschieht mit Hilfe des Graphen G . Nach Satz 1 entsprechen die Linien im TP3 - Schema den einzelnen Variablen und die Punkte in G den einzelnen Gleichungen in (11). Zunächst erhalten die Variablen, die den von z induzierten Endlinien zugeordnet sind, einen willkürlichen Wert. Anschließend wird das Gleichungssystem schrittweise aufgelöst. Dabei ist die einem $P_i \in G$ entsprechende Gleichung von (11) gleichbedeutend mit der Forderung, daß die Summe der zu den drei Linien durch P_i gehörigen Variablen gleich dem c_{ijk} ist, das zu P_i gehört. Deshalb ist eine schrittweise Auflösung von (11) genau dann möglich, wenn es stets mindestens eine solche Gleichung gibt, in der bereits die Werte zweier Variablen bekannt sind.

Satz 4: Z sei eine zyklenfreie Zuordnung der dreieckigen Menge D und G der zugehörige Graph. Dann ist eine sukzessive Ermittlung der Lösung von (11) in der angegebenen Weise anhand G möglich. Genau dann ist dies möglich, wenn z eine zyklenfreie Zuordnung ist.

Damit kann die Lösung von (5) und (6) anhand G gefunden werden, solange die Basis B eine dreieckige Menge bildet. Da in diesem Fall stets $R(\bar{B}) = R(\bar{B}, \bar{a}_{rst})$ gilt, liefert Satz 3 mit der Lösung der Einschränkung auch gleichzeitig die Lösung von (6).

In dem Fall, daß B nicht dreieckig ist, kommt noch ein weiteres Gleichungssystem ins Spiel:

$$u_{jk} + v_{ki} + w_{ij} = 0 \quad \forall ijk \in I_P \quad (12a)$$

$$u_{st} + v_{tr} + w_{rs} = c_{rst} \quad \forall rst \in I_H \quad (12b)$$

Dabei bedeutet (12a) ein System entsprechend (11); die zugehörige Koeffizientenmatrix ist \bar{D}' , wo \bar{D} eine dreieckige Menge bedeutet; (12b) gehört zu einer Menge $\{\bar{a}_{rst} \mid \bar{a}_{rst} \in \bar{A}, rst \in I_H\}$, deren Vektoren sämtliche von D linear unabhängig sowie auch untereinander linear unabhängig sind. Sind dann wieder D, z und G wie bei Satz 2 gegeben, so ist es auch in diesem Fall möglich, anhand G die Lösung von (12) zu finden. Hierzu sind geeigneten Endlinien vorläufige Werte zuzuweisen und dann die Vorgängerbereiche der auf diesen Endlinien liegenden Punkte wieder geeignet zu markieren.

4. Lösung des TP3 bei dreieckiger Basis

Für die Lösung des TP3 gelten grundsätzlich die Ausführungen in 1. mit dem Ablaufschema aus Abb.1. Bildet dann die Basis B eine dreieckige Menge im Sinne der Definition 3, so sind die einzelnen Schritte zur Lösung des TP3 mit Hilfe eines zu B gehörigen Graphen entsprechend Abb. 1 und den Ausführungen in 3. durchführbar.

Zunächst ist eine zu B gehörende zyklenfreie Zuordnung z zu ermitteln und der zugehörige Graph. Sei dies z.B. der in Abb.3 angegebene Graph. Die Schattenpreise erhält man dann anhand dieses Graphen entsprechend Satz 4: Die zu den Endlinien gehörigen Schattenpreise erhalten einen beliebigen Wert, zweckmäßigerweise 0. In Abb.5 ist dies durchgeführt: sofern eine Endlinie einer Horizontalen bzw. Vertikalen im TP3 - Schema entspricht, steht der zugehörige Schattenpreis 0 rechts bzw. unten am Rand des Schemas; bei den Endlinien, die einer Diagonalen in einem der Quadrate entsprechen, steht der Schattenpreis in der linken unteren Ecke des betreffenden Quadrates. Abb.5 läßt erkennen, daß in all den Gleichungen, die

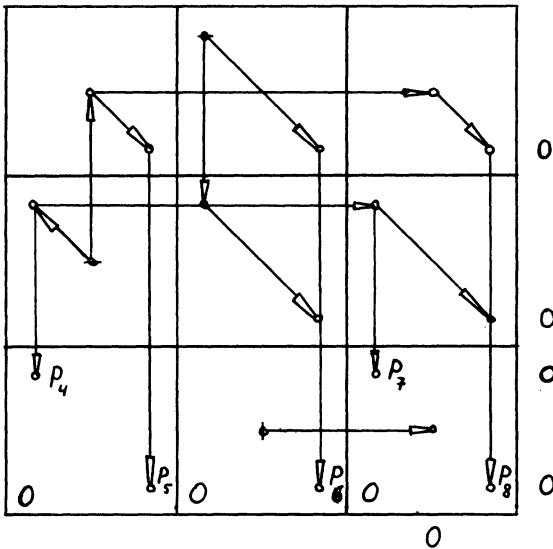


Abb.5: Ermittlung der Schattenpreise

den Punkten P_4 , P_5 , P_6 , P_7 , P_8 entsprechen, dann jeweils 2 Schattenpreise bekannt sind und damit der jeweilige dritte vollends ermittelt werden kann. Satz 4 garantiert, daß dieser Prozeß sukzessive bis zur vollständigen Ermittlung sämtlicher Schattenpreise durchführbar ist. Für die reduzierten Kosten sind für jede Position im TP3 - Schema vom zugehörigen Preis c_{ijk} (aus (2)) die Schattenpreise der drei durch (i, j, k) gehenden Linien zu subtrahieren. Zweckmäßigerweise erledigt man dies gleich nachdem ein Schattenpreis gefunden wurde, indem man ihn bei sämtlichen Positionen, die auf der zugehörigen Linie liegen, subtrahiert.

Sind nicht sämtliche reduzierten Kosten ≥ 0 , so wählt man eine der Nichtbasisvariablen (NBV) zur Aufnahme in die Basis. Als Kriterium kann z.B. der kleinste Wert der \hat{c}_{ijk} dienen. Sei dies z.B. die zu $(2,2,2)$ gehörige NBV, Man braucht dann für Schritt (6) die Lösung von

$$\bar{B} \bar{y} - \bar{a}_{222} ;$$

sie wird anhand G nach Satz 3 gefunden. In Abb. 4 wurde dies bereits durchgeführt. Die Werte von \bar{y} stehen dann bei den den je-

weiligen Y_{ijk} entsprechenden Punkten P_i . Damit kann Schritt (7) durchgeführt werden; man benötigt hierzu noch die Werte der Basisvariablen (BV). Der Austauschschritt selbst sieht dann formal so aus, daß die aufzunehmende NBV einen in Schritt 7 ermittelten Wert θ erhält und die Werte der einzelnen BV sich um $Y_{ijk} \cdot \theta$ ändern; alle hierzu benötigten Größen sind bereits bekannt.

Damit ist ein Schritt der Iteration durchgeführt. Für die neue Basis ist nun zuerst wieder ein neuer Graph zu ermitteln; in vielen Fällen erhält man ihn durch eine geeignete Modifikation des vorliegenden Graphen. Dann kann der nächste Iterations - schritt beginnen.

5. Lösung des TP3 bei nichtdreieckiger Basis

Es ist keineswegs gesichert, daß bei diesen Iterationen die Basis B stets dreieckig im Sinne der Def.3 bleibt. Ist sie nicht mehr dreieckig, so versagt das vorgeschlagene Verfahren zunächst, da es nun keinen zyklensfreien Graphen zu B mehr gibt, dies aber notwendige Voraussetzung für die Lösung nach Satz 3 und 4 ist. Bei dem in 4. beschriebenen Vorgehen ist deshalb bei den Schritten 2 und 6 für diesen Fall ein neuer Weg zu finden. Hierzu entfernt man zunächst aus B solange Vektoren, bis die verbleibende Menge $D \subset B$ wieder dreieckig ist. Gewöhnlich tritt dieser Fall rasch ein; bei durchgerechneten größeren Beispielen mußten maximal 4 Vektoren entfernt werden. Sei

$$B = D \cup H \quad (13)$$

eine dementsprechende Zerlegung von B in zwei disjunkte Mengen, wo

$$D = \{ \bar{a}_{ijk} \mid \bar{a}_{ijk} \in B, ijk \in I_D \} \quad (14a)$$

eine dreieckige Menge und

$$H = \{ \bar{a}_{ijk} \mid \bar{a}_{ijk} \in B, ijk \in I_H \} \quad (14b)$$

bedeuten. Für Schritt 6 ist die Lösung von

$$\bar{B} \bar{y} = - \bar{a}_{rst} \quad (15)$$

gesucht. Anstatt (15) wird zunächst

$$\bar{D} \bar{y}^D = - \bar{a}_{rst} \quad (16)$$

betrachtet. Da D dreieckig ist, gibt es hierzu sicher ein z mit einem zugehörigen zyklensfreien Graphen G , anhand dem nach Satz 3 die Lösung der Einschränkung von (16) auf G gefunden werden kann. Die dabei erhaltenen Werte

$$y_{ijk}^D \quad ijk \in I_D$$

sind jedoch noch keine Lösung von (15). Für eine solche macht man den Ansatz

$$y_{ijk} = y_{ijk}^D + \sum_{uvw \in I_H} x_{ijk}^{uvw} \quad \forall ijk \in I_D \quad (17a)$$

$$y_{uvw} = y_{uvw} \quad \forall uvw \in I_H \quad (17b).$$

Zur Ermittlung der unbekannten Größen in (17) setzt man (17) in (15) ein. Dies führt auf $n_H = |I_H|$ Gleichungssysteme der Form (16) für die Variablen x_{ijk}^{uvw} / y_{uvw} , die wieder anhand des vorliegenden zyklensfreien Graphen G gelöst werden können. Damit bleiben noch die y_{uvw} als Unbekannte übrig. Zu ihrer Ermittlung wählt man n_H geeignete Endlinien aus, anhand denen lineare Bestimmungsgleichungen für die y_{uvw} aufgestellt werden können. Damit bleibt schließlich noch die Lösung von n_H linearen Gleichungen in n_H Unbekannten übrig, was aber angesichts der geringen Größe von n_H mühelos mit den üblichen Methoden zu bewältigen ist. Dann können nach (17a) die restlichen y_{ijk} vollends berechnet werden und die Lösung von (15) ist gefunden. Schritt 6 ist damit erledigt.

Für Schritt 2 ist die Lösung von

$$u_{jk} + v_{ki} + w_{ij} = c_{ijk} \quad \forall ijk \in I_B \quad (18)$$

gesucht. Zum zyklensfreien Graphen G von D gehört das Teilsystem von (18) mit $ijk \in I_D$; es ist gleichwertig mit (11) und seine Lösung ist entsprechend Satz 4 möglich. Sei dies

$$u_{jk}^D, v_{ki}^D, w_{ij}^D \quad \text{mit } ijk \in I_D.$$

Um hieraus die Lösung von (18) zu erhalten, setzt man

$$\begin{aligned} u_{jk} &= u_{jk}^D + du_{jk} \\ v_{ki} &= v_{ki}^D + dv_{ki} \\ w_{ij} &= w_{ij}^D + dw_{ij} \end{aligned} \quad (19)$$

(19) in (18) eingesetzt führt auf ein Gleichungssystem für die du_{jk} , dv_{ki} , dw_{ij} , das gerade die Form (12) besitzt. Nach den Ausführungen in 3. kann dieses ebenfalls anhand des zu D gehörigen zyklischen Graphen G gelöst werden; damit erhält man die du_{jk} , dv_{ki} , dw_{ij} und damit nach (19) die gesuchten Schattenpreise von (18). Damit ist auch Schritt 2 für eine nichtdreieckige Basis bewältigt.

Die übrigen Schritte verlaufen wie beim Fall einer dreieckigen Basis angegeben. Damit ist es grundsätzlich möglich, mit Hilfe des in 2. entwickelten Kalküls in jedem Fall das TP3 zu lösen. Für die Durchführung der Rechnung mit Hilfe einer Rechenanlage ist noch von Bedeutung, daß für G kein Speicherplatz benötigt wird, sofern z gespeichert ist. Für die Speicherung von z wiederum genügen $mn + np + pm$ Speicherplätze, wenn $m = I$, $n = J$ und $p = K$ bedeuten. Insgesamt gestattet das Verfahren eine äußerst komprimierte Speicherung der benötigten Daten. Die Rechenzeit für einen Iterationsschritt hängt davon ab, ob die jeweilige Basis dreieckig ist oder nicht. Bei nichtdreieckiger Basis ist der Rechenaufwand größer als bei dreieckiger Basis und ungefähr proportional der Zahl n_H der aus der Basis entfernten Vektoren.

Literatur

- [1] BARTH, U.: Das dreidimensionale Transportproblem.
Forschungsbericht, Mainz, 1968.
- [2] HALEY, K.B.: The Solid Transportation Problem.
JORSA 10 (1962), 448 - 463.
- [3] HALEY, K.B.: The Multi-Index Problem.
JORSA 11 (1963), 368- 379.
- [4] JUNGINGER, W.: Über die Lösung des dreidimensionalen
Transportproblems. Dissertation,
Universität Stuttgart, 1971.
- [5] MÜLLER-MERBACH, H.: Die Lösung des Transportproblems
auf Rechenanlagen - Ein ALGOL-Programm.
Elektronische Datenverarbeitung 8 (1966),
49 - 56.

Zur multiparametrischen linearen Programmierung

von T. Gal, Aachen

Unter multiparametrischer linearer Programmierung wird ein parametrisches lineares Programm verstanden mit einem Vektorparameter in der rechten Seite, oder in der Zielfunktion, oder in der Restriktionsmatrix.

Für die multiparametrischen Fälle:

$$1) \max z = c^T(v)x, \text{ so daß } Ax = b, x \geq 0, \text{ mit } c(v) = c + H_v,$$

$$2) \max z = c^T x, \text{ so daß } Ax = b(\lambda), x \geq 0, \text{ mit } b(\lambda) = b + F\lambda,$$

wo $c \in E^n$, $x \in E^n$, $b \in E^m$, A eine (m, n) , H eine (n, S) , F eine (m, s) Matrix, und $v \in E^S$, $\lambda \in E^S$ Vektorparameter sind,

wurde ein allgemeiner Lösungsalgorithmus ausgearbeitet (Gal, Nedoma 1972), der auf einem graphentheoretischen Verfahren beruht (Manas, Nedoma 1968).

In diesem Beitrag werden kurz zwei weitere Fälle behandelt.

I. Bezeichnungen

Es sei $J = \{1, \dots, n\}$ die Indexmenge aller Variablen, $\rho = \{j_1, \dots, j_m\}$ eine Indexmenge von Basisvariablen. Die Menge ρ wird Basisindex genannt. Da im allgemeinen ein parametrisches Problem mehrere optimale Basen haben kann, werden die Basisindizes durch Indizes unterschieden, z.B. ρ_u .

Durch B_ρ , bzw. B_ρ^{-1} wird die Basismatrix (kurz: Basis), bzw. ihre Inverse bezeichnet (für ρ_u steht B_u bzw. B_u^{-1}). Dann gilt:

$$\rho_F = B_\rho^{-1}F, \quad c_B^T = (c_{j_1}, \dots, c_{j_m}),$$

$$\rho_c^T = c_B^T B_\rho^{-1}A - c^T \text{ mit den Elementen } \rho_\Delta c_j = \rho c_j - c_j = c_B^T \rho a^j - c_j,$$

$$\rho a^j = B_\rho^{-1}a^j, \quad a^j \text{ die } j\text{-te Spalte der Matrix } A,$$

$$z^{(\rho)} = c_B^T \rho b \quad \text{mit} \quad \rho b = B_\rho^{-1}b.$$

Ist in der Maximierungsaufgabe

$$\rho b \geq 0 \quad \text{und} \quad \rho c \geq 0,$$

wird die Basis B_ρ optimale Basis genannt.

II. Homogene multiparametrische lineare Programmierung ^{*})

Es sei das folgende Problem gegeben: Ermittle einen Bereich $K \subset E^S$, so daß die Aufgabe: Maximiere

$$(II.1) \quad z = c^T x$$

unter den Nebenbedingungen

$$(II.2) \quad Ax = F\lambda, \quad x \geq 0,$$

für alle $\lambda \in K$ eine endliche nichttriviale optimale Lösung bez. x hat und für $\lambda \in E^S - K$ keine (nichttriviale) optimale Lösung bez. x besitzt.

Man setze voraus, daß es mindestens einen optimalen Basisindex gibt, so daß die Aufgabe (II.1), (II.2) eine endliche nichttriviale optimale Lösung hat. Dann ist durch

$$(II.3) \quad B_p^{-1} F\lambda \geq 0 \quad \text{oder} \quad -\rho_p \lambda \leq 0$$

eindeutig ein kritischer Bereich $R_p \subset E^S$, $R_p \subset K$, definiert, so daß für, und nur für $\lambda \in R_p$ die Basis B_p optimal ist.

Der gegebenen Aufgabe wird ein zusammenhängender nichtgerichteter Graph zugeordnet (Einzelheiten vgl. Gal, Nedoma, 1972, oder Gal 1972a), wobei jeder Knoten ρ des Graphen ein optimaler Basisindex der Aufgabe (II.1), (II.2) ist. Zwei Knoten ρ_1, ρ_2 sind durch eine Kante verbunden, wenn es möglich ist von B_1 zu B_2 und umgekehrt ~~an~~ einem dualen Simplexschritt überzugehen. Zwei Basen B_1, B_2 , die die erwähnte Eigenschaft haben und für die ein $\tilde{\lambda} \in K$ so existiert, daß für $\lambda = \tilde{\lambda}$ beide Basen zugleich optimale Basen sind, werden Nachbarbasen genannt. Die entsprechenden Knoten bzw. kritischen Bereiche werden demgemäß Nachbarknoten bzw. Nachbarbereiche genannt.

II.1. Das Lösungsverfahren

Das gegebene Problem wird in zwei Phasen gelöst.

Phase 1:

a) Bestimme eine nichttriviale zulässige Lösung (x^0, λ^0) , $x^0 \geq 0$, $\lambda^0 \neq 0$, des Systems

^{*}) Erscheint demnächst in allen Einzelheiten in der Zeitschrift Unternehmensforschung.

$$(II.4) \quad Ax - F\lambda = 0, \quad x \geq 0.$$

b) Bestimme ~~die~~ optimale Lösung der Aufgabe

$$(II.5) \quad \max z = c^T x, \text{ so da\ss } Ax = F\lambda^0, \quad x \geq 0.$$

Phase 2:

Bestimme K , ausgehend von dem in Phase 1 bereits ermittelten kritischen Bereich R_0 .

Die Phase 1 wird grundsätzlich wie folgt durchgeführt:

Man setze $\lambda_1^+ = \lambda_2^+ = \dots = \lambda_s^+ = 1$ und man löse das System

$$(II.6) \quad Ax = F\lambda^+, \quad x \geq 0.$$

Existiert keine zulässige Lösung des Systems (II.6) bez. x , so bezieht man schrittweise die Parameter λ_k in die Lösung ein, wobei die Bedingung $\lambda_k^+ = 1$ aufgehoben wird.

Ist auf diese Weise in endlich vielen primalen Simplexschritten eine zulässige Lösung (x^0, λ^0) , $x^0 \geq 0$, $\lambda^0 \neq 0$, ermittelt, so löst man die Aufgabe (II.5). Die optimale Lösung dieser Aufgabe (falls sie existiert) ergibt dann den optimalen Basisindex ρ_0 , damit B_0 und R_0 bekannt, woraus $K \neq \emptyset$ folgt.

Die Phase 2 läuft grundsätzlich nach dem Algorithmus aus (Gal, Nedoma 1972 oder Gal 1972a).

Auf ähnliche Weise wird auch das nachstehende Problem gelöst: Ermittle einen Bereich $K \subset E^S$, so daß die Aufgabe: Maximiere

$$z = (Hv)^T x,$$

unter den Nebenbedingungen

$$Ax = b, \quad x \geq 0,$$

für alle $v \in K$ eine endliche nichttriviale optimale Lösung bez. x hat und für $v \in E^S - K$ keine optimale nichttriviale Lösung bez. x existiert.

Als Sonderfall des Problems (II.1), (II.2) kann die folgende Aufgabe gelöst werden:

$$\max z = c^T x, \text{ so da\ss } Ax = F_D \lambda, \quad x \geq 0, \quad \lambda \geq 0,$$

wo

$$F_D = \begin{pmatrix} b_1, & \dots, & 0 \\ 1 & \dots & \dots \\ 0, & \dots, & b_m \end{pmatrix} \quad \text{und } \lambda \in E^m \text{ ist.}$$

Die erste Phase des Lösungsverfahrens bleibt unverändert, in der zweiten Phase muß jede ermittelte optimale Basis getestet werden, ob der zuständige kritische Bereich einen nichtleeren Durchschnitt mit $\lambda \geq 0$ hat. Dazu wird jeweils eine Simplex-Aufgabe gelöst, die Eingehend in (Gal, 1972b) beschrieben ist.

III. Doppelt-multiparametrische lineare Programmierung mit abhängigen Parametern

Es sei das folgende Problem gegeben: Ermittle den Definitionsbereich $\tilde{K} \subset E^S$ der linearen Funktion

$$(III.1) \quad v = f(\lambda), \quad v \in E^S, \quad \lambda \in E^S,$$

so daß die Aufgabe: Maximiere

$$(III.2) \quad z(v, \lambda) = c^T(v, \lambda)x$$

unter den Nebenbedingungen

$$(III.3) \quad Ax = b(v, \lambda), \quad x \geq 0,$$

für alle $\lambda \in \tilde{K}$ eine endliche optimale Lösung bez. x hat und für $\lambda \in E^S - \tilde{K}$ keine optimale Lösung existiert.

Dabei ist

$$(III.4) \quad f(\lambda) = d + D\lambda,$$

$$(III.5) \quad c(v, \lambda) = c + H^1 v + F^1 \lambda, \quad b(v, \lambda) = b + H^2 v + F^2 \lambda$$

mit festen Vektoren und Matrizen: H^1 eine (N, S) , F^1 eine (N, s) , D eine (S, s) , H^2 eine (m, S) , F^2 eine (m, s) , $d \in E^S$, A eine (m, N) , $c \in E^N$, $b \in E^m$, $x \in E^N$.

Durch Einsetzen von (III.4), (III.5) in (III.2) und (III.3) erhält man die Aufgabe

$$(III.6) \quad \max \quad z(\lambda) = (c^* + F^{*1} \lambda)^T x,$$

so daß

$$(III.7) \quad Ax = b^* + F^{*2} \lambda, \quad x \geq 0,$$

wo

$$(III.8) \quad c^* = c + H^1 d, \quad F^{*1} = H^1 D + F^1, \quad b^* = b + H^2 d, \quad F^{*2} = H^2 D + F^2.$$

Man setze voraus, daß $\lambda \in E^S$, und somit $\tilde{v} \in E^S$, existiert, so daß die Aufgabe (III.6), (III.7) eine endliche optimale Lösung bez. x hat. Damit ist durch

$$(III.9) \quad -\rho F^{*1} \lambda \leq \rho c^*$$

ein kritischer Bereich \bar{R}_ρ , und durch

$$(III.10) \quad -\rho F^{*2} \lambda \leq \rho b^*$$

ein kritischer Bereich R_ρ definiert. Die Basis B_ρ bleibt optimal für alle $\lambda \in R_\rho^*$, wobei

$$(III.11) \quad R_\rho^* = \bar{R}_\rho \cap R_\rho.$$

III.1. Das Lösungsverfahren ^{x)}

Da auch dieser Fall in einer Veröffentlichung eingehend behandelt wird, beschränken wir uns hier nur auf die Prinzipien.

^{x)} Der Aufsatz ist vor einigen Tagen eingereicht worden. Eine eingehende Behandlung des diskutierten Falles ist auch in (Gal, 1972a) zu finden.

Das Optimum einer Zielfunktion ist nur über eine nichtleere Lösungsmenge sinnvoll zu ermitteln. Deshalb wird in der

Phase 1

zunächst auf λ in $c(\lambda)$ verzichtet und man ermittelt eine Lösung (x^0, λ) des Systems

$$(III.12) \quad Ax - F^{*2}\lambda = b^*, \quad x \geq 0.$$

Falls (x^0, λ) existiert, löst man die Aufgabe:

$$(III.13) \quad \max \quad z = (c^* + F^{*1}\lambda)^T x$$

so daß

$$(III.14) \quad Ax = b^* + F^{*2}\lambda, \quad x \geq 0.$$

Existiert eine dual und primal zulässige Lösung (d.h. eine optimale Lösung) der Aufgabe (III.13), (III.14), so geht man zur Phase 2 über. Falls für $\lambda = \lambda$ keine dual zulässige Lösung der Aufgabe (III.13), (III.14) existiert, so verzichtet man wieder auf λ in $c(\lambda)$ und ermittelt sukzessive die Nachbarbasen für den Fall mit einem Vektorparameter in der rechten Seite. In jeder der ermittelten Basen bestimmt man R_ρ und \bar{R}_ρ und stellt fest, ob $R_\rho^* = R_\rho \cap \bar{R}_\rho \neq \emptyset$. Ist dies für ein bestimmtes ρ der Fall, so ist dadurch eine optimale Lösung des ursprünglichen Problems gefunden.

Phase 2:

Man ermittelt sukzessive die Nachbarbasen gemäß eines modifizierten Algorithmus aus (Gal, Nedoma 1972). In jeder dieser Basen wird festgestellt, ob $R_\rho^* \neq \emptyset$. Ist dies der Fall, wird der Basisindex gespeichert. Anderenfalls wird dieser Basisindex in einen neuen Speicher eingeführt, so daß diese Basis im Laufe der Berechnungen nicht mehr in Frage kommt.

Im Vergleich mit den anderen multiparametrischen Fällen, bei denen der Gesamtbereich K eine konvexe Menge bildet, ist der Bereich \tilde{K} für den doppelt-multiparametrischen Fall im allgemeinen nicht konvex.

Literaturverzeichnis

Gal, T., Nedoma, J.: Multiparametric linear Programming.

Management Sci., Theory Series, March Issue 1972

Gal, T.: Betriebliche Entscheidungen und parametrische Programmierung. de Gruyter Verlag Berlin, New York 1972a

Gal, T.: Homogene mehrparametrische lineare Programmierung.

Unternehmensforschung 1972b

Mañas, M., Nedoma, J.: Finding all vertices of convex polyhedron.

Numer. Math. 12, 1968, S. 226 - 229

**Zur Lösung eines Zwei-Stufen-Risiko Modells
der stochastischen linearen Optimierung**
von W. Bühler, Aachen

Zusammenfassung

Eine Reformulierung eines stochastischen linearen Programmierungsproblems führt auf ein nichtlineares deterministisches Ersatzproblem der Form

$$\min_{x \in K} \mu(x) + \sqrt{G(x)},$$

wobei $\mu(x)$ der Erwartungswert und $G(x)$ die Standardabweichung einer von $x \in K$ abhängigen Zufallsvariablen ist. Für diese Aufgabe wird ein Lösungsverfahren angegeben.

1. Formulierung und Eigenschaften des Modells

Gegeben sei das Problem

$$\begin{aligned} \min \quad & C^T x \\ \text{s.t.} \quad & Ax \leq B \\ & x \in K \end{aligned} \quad (1),$$

wobei (B, C) eine $m+n$ dimensionale reelle Zufallsvariable über einem Maßraum (Ω, \mathcal{U}, P) , $K \neq \emptyset$ ein kompaktes, konvexes Polyeder und A eine $m \cdot n$ Matrix ist.

In dieser Form ist (1) keine sinnvoll gestellte Optimierungsaufgabe. Als Interpretation von (1) wird daher das folgende nichtlineare deterministische Ersatzproblem betrachtet:

$$\max_{x \in K} P\{(b, c) \in \Omega : c^T x + Q(b - Ax) \leq v\} =: \max_{x \in K} F_x(v) \quad (2),$$

wobei $Q(B - Ax) = \sum_{i=1}^m Q_i((B - Ax)_i)$ für alle $x \in K$ eine $(\mathcal{B}^m, \mathcal{B}^1)$ -

meßbare Strafkostenfunktion ^{1.)} und v eine fest vorgegebene reelle Zahl ist.

(2) enthält Elemente des Zwei - Stufen - Modells insofern als eine mögliche Verletzung der Nebenbedingungen $Ax \leq B$ nach Wahl eines Entscheidungsvektors x und nach Realisation b der Zufallsgröße B kompensiert werden kann, wobei die Kosten $Q(b - Ax)$ entstehen.

Die Zielfunktion von (2) entspricht der Zielfunktion eines speziellen Chance - Constrained Problems, dem P - Modell (vgl. [C]). Es wird dabei die volle in der Verteilungsfunktion F_x von $C^T x + Q(B - Ax)$ steckende Information berücksichtigt und nicht nur, wie beim Zwei - Stufen - Modell, der Erwartungswert (vgl. [K]).

An anderer Stelle (vgl. [B]) wurden Bedingungen angegeben, unter denen (2) durch das Optimierungsmodell

$$\min_{x \in K} \mu(x) + \lambda \cdot \sigma(x) \quad (A)$$

ersetzt werden kann, wobei $\mu(x)$ Erwartungswert, $\sigma(x)$ Standardabweichung von $C^T x + Q(B - Ax)$ und $\lambda \geq 0$ ist. Ebenfalls in [B] wurden Beweisskizzen für die im folgenden benötigten Aussagen gegeben.

A 1: Sind B, C und die Komponenten von B stochastisch unabhängig, dann existiert ein zu

$$\min_{x \in K} \mu(x) + \frac{\lambda}{2R} \sigma^2(x) \quad H_R \quad (0 \leq R \leq \infty)$$

1.) \mathcal{B}^k bezeichne die Borel algebra des \mathbb{R}^k .

äquivalentes, separables Optimierungsproblem mit linearen Nebenbedingungen. Dabei ist H_0 als

$$\min_{x \in K} G^2(x) \quad \text{und } H_\infty \text{ als}$$

$$\min_{x \in K} \mu(x)$$

definiert.

A 2: Ist die Zielfunktion von H_R streng quasikonvex und differenzierbar, dann existiert ein \bar{R} und eine optimale Lösung $\hat{x}(\bar{R})$ von $H_{\bar{R}}$, so daß $\hat{x}(\bar{R})$ optimale Lösung von (A) ist und

$$\bar{R} = G(\hat{x}(\bar{R})) \quad 1.)$$

gilt, d.h. \bar{R} ist Fixpunkt von $G(\hat{x}(R))$.

A 3: Ist $\mu(x)$ und $G(x)$ stetig in x und F_x nicht degeneriert für alle $x \in K$, dann ist

a.) $G(\hat{x}(R))$ monoton wachsend in R

b.) $0 < G(\hat{x}(0)) \leq G(\hat{x}(\infty)) < \infty$.

Schließlich wird noch folgende Definition benötigt:

Def.: $\bar{R} \in \mathbb{R}^+$ heißt schwacher Fixpunkt, falls
für alle $\varepsilon > 0$ ein $R \in \mathbb{R}^+$ existiert mit
 $|R - \bar{R}| < \varepsilon$ und $|G(\hat{x}(R)) - \bar{R}| < \varepsilon$

Aus dieser Definition ergibt sich unmittelbar, daß jeder Fixpunkt ein schwacher Fixpunkt ist.

1.) $\hat{x}(R)$ wird hier als Abbildung des \mathbb{R}^+ in den \mathbb{R}^n aufgefaßt, die jedem Element $R \in \mathbb{R}^+$ ein $\hat{x}(R)$ aus der optimalen Lösungsmenge $X^*(R)$ von H_R zuordnet. (Bildelement und Abbildung erhalten dieselbe Bezeichnung, es werden sich keine Verwechslungen ergeben). Ist für ein $R \in \mathbb{R}^+$ die Menge $X^*(R)$ mehrelementig, dann existieren verschiedene derartige "optimale Abbildungen $\hat{x}(R)$ ". Die folgenden Aussagen gelten für irgendein $\hat{x}(R)$.

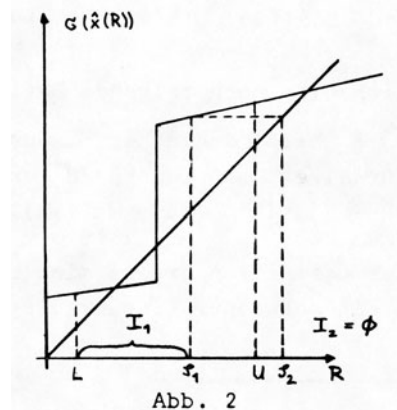
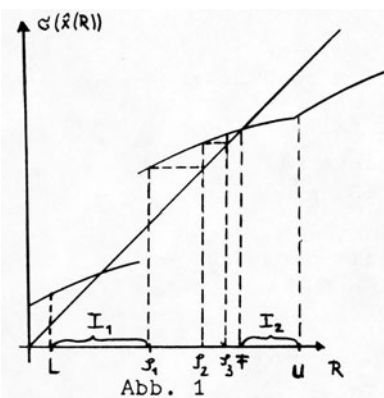
2. Qualitative Beschreibung eines Verfahrens zur Lösung des Zwei - Stufen - Risiko Modells

Auf A2 und A3 beruht ein Bisektionsverfahren zur Bestimmung sämtlicher Fixpunkte von $G(\hat{x}(R))$ und damit aller optimalen Lösungen von (A).

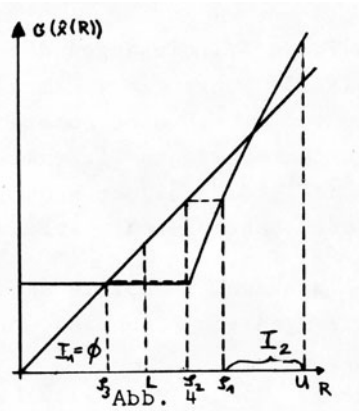
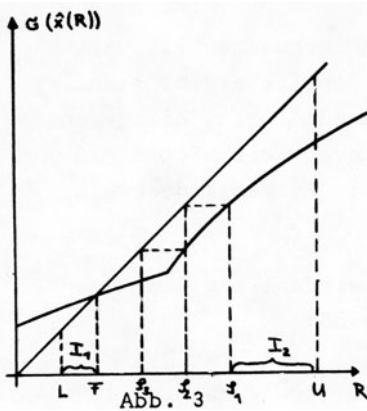
Zur Beschreibung des Verfahrens sei zunächst unterstellt, daß das abgeschlossene Intervall $I = [L, U]$ ($0 < L < U < \infty$) alle noch nicht berechneten Fixpunkte enthält. Ferner sei \mathcal{J}_1 aus dem Innern von I (z.B. $\mathcal{J}_1 = \frac{L+U}{2}$) ein Startwert für die folgende Iterationsvorschrift

$$\text{Löse } H_{\mathcal{J}_i} \Rightarrow \hat{x}(\mathcal{J}_i) \Rightarrow \mathcal{J}_{i+1} := G(\hat{x}(\mathcal{J}_i)) \quad i = 1, 2, \dots \quad (\text{IT})$$

Ist $G(\hat{x}(\mathcal{J}_1)) > \mathcal{J}_1$, dann ist die durch (IT) definierte Folge \mathcal{J}_i monoton wachsend und konvergiert entweder gegen einen schwachen Fixpunkt $F \leq U$ (Abb. 1) oder es existiert ein $i^* > 1$, so daß $\mathcal{J}_{i^*} > U$ ist (Abb. 2).



Ist $G(\hat{x}(\mathcal{J}_1)) < \mathcal{J}_1$, dann ist die Folge \mathcal{J}_i monoton fallend und konvergiert entweder gegen einen schwachen Fixpunkt $F > L$ (Abb. 3) oder es existiert ein $i^* > 1$, so daß $\mathcal{J}_{i^*} < L$ (Abb. 4).



Der Iterationsprozeß wird abgebrochen, falls $s_i \notin [L, U]$ für ein $i \in \mathbb{N}$ oder falls eine Verletzung der strengen Monotonie in der Folge s_i auftritt. Im zweiten Fall wird der vorletzte Iterationswert als Fixpunkt akzeptiert.

Ist $G(\hat{x}(s_1)) = s_1$, dann endet der Iterationsprozeß (IT) nach einem Schritt.

In jedem der möglichen Fälle kann die weitere Suche nach Fixpunkten auf die Intervalle I_1 und I_2 (vgl. Abb. 1 bis 4) beschränkt werden, wobei I_1 oder I_2 leer sein kann.

Falls $I_2 \neq \emptyset$ wird die Suche nach weiteren Fixpunkten in I_2 fortgesetzt, andernfalls in I_1 . Als Ergebnis des erneut anzuwendenden Iterationsprozesses (IT) erhält man wieder zwei Teilintervalle I_{21} und I_{22} bzw. I_{11} und I_{12} (wobei eines leer sein kann), die für die weitere Fixpunktsuche in Frage kommen.

Ist nach k Schritten das Verfahren noch nicht beendet, dann sind höchstens 2^k Teilintervalle von I konstruiert, die möglicherweise noch Fixpunkte enthalten. Die Suche wird nun in dem Intervall mit dem lexikographisch größten Indexvektor fortgesetzt. Dabei wird ein Indexvektor mit Dimension 1 ($1 \leq k$) stets als lexikographisch kleiner als ein solcher mit Dimension m ($m \leq k$) definiert, falls $1 < m$ ist.

Das Verfahren wird gestartet mit dem Intervall $\bar{I} = [0, \infty]$ und dem ersten Startwert $s_1 = \infty$. Man erhält dann eine monoton

fallende Folge \mathcal{V}_i , die gegen den größten schwachen Fixpunkt F' von $\mathcal{G}(\hat{x}(R))$ konvergiert. Im zweiten Schritt ergibt sich ausgehend von $\mathcal{V}_1 = 0$ eine monoton wachsende Folge, die gegen den kleinsten schwachen Fixpunkt F'' konvergiert. Setzt man $L := F''$, $U := F'$, dann eignet sich $[L, U]$ als Ausgangsintervall für das oben beschriebene Verfahren.

Es sollen nun zwei Beispiele angegeben werden, für die die Voraussetzungen von A2 und A3 erfüllt sind:

$$\text{Ist } Q(B-Ax) = \sum_{i=1}^m q_i ((B-Ax)_i^+)^2 \quad (q_i > 0, \quad y^+ = \max(0, y))$$

und besitzt das Randmaß P_i von B_i in jedem kompakten Intervall höchstens endlich viele Unstetigkeitsstellen, dann ist H_R konvex, während dies für (A) nicht zu gelten braucht. Außerdem ist in diesem Fall jeder schwache Fixpunkt auch Fixpunkt.

Ist $Q(B-Ax) = \sum_{i=1}^m q_i ((B-Ax)_i)^2$, dann ist H_R ein konvexes quadratisches Optimierungsproblem. Ferner läßt sich zeigen, daß die Menge der Fixpunkte von $\mathcal{G}(\hat{x}(R))$ zusammenhängend ist, wodurch sich das oben beschriebene Verfahren wesentlich vereinfacht, da lediglich die Berechnung von F' und F'' erforderlich ist.

3. Anwendung auf ein Investitionsproblem

Als Beispiel für das oben dargestellte Verfahren wird ein auf Massé und Gibrat zurückgehendes Investitionsproblem gewählt, das zur Planung von Erweiterungsinvestitionen im Erzeugungsbereich der Electricité de France aufgestellt wurde ^{1.)}. Es stehen dabei zur Deckung der durch die Größen "Garantierte Leistung B_1 " ($B_1 = 1692$ MW), "Spitzenleistung B_2 " ($B_2 = 2307$ MW) und "Jahreserzeugung B_3 " ($B_3 = 7200$ GWh) charakterisierte Nachfrage nach Energie fünf verschiedene Kraftwerktypen zur Verfügung:

1.) Für Einzelheiten vgl. [M], S. 108 ff.

1. Dampfkraftwerke (x_1)
2. Laufwasserkraftwerke (x_2)
3. Wasserkraftwerke mit großem Speicher (x_3)
4. Wasserkraftwerke mit kleinem Speicher (x_4)
5. Gezeitenkraftwerke (x_5)

Das Investitionsbudget beträgt 2 710 Mio N.F. Als Planungsziel wird die Minimierung der auf einen festen Zeitpunkt diskontierten Investitions- und Betriebsausgaben unterstellt.

Es ergibt sich dann das folgende lineare Modell:

$$\begin{aligned}
 \text{Min} \quad & 1,36x_1 + 0,56x_2 + 1,01x_3 + 1,04x_4 + 0,79x_5 \\
 & x_1 + x_2 + x_3 + x_4 + x_5 \geq 1692 \\
 & 1,15x_1 + 1,10x_2 + 1,20x_3 + 3,00x_4 + 2,13x_5 \geq 2307 \quad (3) \\
 & 7,00x_1 + 12,60x_2 + 1,30x_3 + 7,35x_4 + 5,47x_5 \geq 7200 \\
 & 0,97x_1 + 4,20x_2 + 1,30x_3 + 3,10x_4 + 2,13x_5 \leq 2710 \\
 & x \geq 0
 \end{aligned}$$

Im folgenden wird nun angenommen, daß die Nachfragegrößen B_i ($i=1,2,3$) unabhängig normalverteilt sind mit einer Standardabweichung, die etwa 10% des Erwartungswertes beträgt. Es wird also gefordert, daß B_1 (1692, 170); B_2 (2307, 230) und B_3 (7200, 720) normalverteilt sind. Bezüglich der Zielfunktionskoeffizienten C_j werden keine Verteilungsannahmen getroffen. Lediglich Erwartungswert $E(C)$ und Kovarianzmatrix V sind durch

$$E(C^t) = (1,36; 0,56; 1,01; 1,04; 0,79)$$

$$V = \begin{pmatrix} 0,48 & 0,18 & 0 & 0 & 0 \\ 0 & 0,3 & 0,33 & 0,24 & 0 \end{pmatrix}$$

vorgegeben.

Das konvexe, kompakte Polyeder K wird durch die einzige deterministische Restriktion

$$0,97x_1 + 4,20x_2 + 1,30x_3 + 3,10x_4 + 2,13x_5 \leq 2710$$

und die Nichtnegativitätsforderung $x \geq 0$ definiert.

Als Strafkostenfunktion wird eine stückweise lineare Funktion der Form

$$Q(B-Ax) = \sum_{i=1}^3 q_i (B-Ax)_i^+ = \sum_{i=1}^3 q_i s_i ((B-E(B)+E(B)-Ax)_i^+) / s_i \\ =: \sum_{i=1}^3 q_i s_i (\bar{B}-z)_i^+$$

unterstellt, wobei s_i die Standardabweichung von B_i ist $q_1 = 1,538$; $q_2 = 0$; $q_3 = 0,036$ gleich den Optimalwerten der Dualvariablen von (3) gesetzt wird.

Für Erwartungswert $\mu_i(z_i)$ und Varianz $G_i^2(z_i)$ von $q_i s_i (\bar{B}-z)_i^+$ ergibt sich

$$\mu_i(z_i) = q_i s_i \left\{ \frac{1}{\sqrt{2\pi}} e^{-1/2 z_i^2} - z_i (1 - \Phi(z_i)) \right\} \\ G_i^2(z_i) = q_i^2 s_i^2 \left\{ z_i^2 \Phi(-z_i) \Phi(z_i) - \frac{z_i}{\sqrt{2\pi}} e^{1/2 z_i^2} \cdot (1 - 2\Phi(z_i)) + \Phi(-z_i) - \frac{1}{2\pi} e^{-z_i^2} \right\},$$

wobei Φ die Verteilungsfunktion einer (0,1) normalverteilten Zufallsgröße ist.

Das Gesamtmodell lautet damit:

$$\text{Min } E(C^t)x + \sum_{i=1}^3 \mu_i(z_i) + \mathcal{J} \cdot \sqrt{x^t V x + \sum_{i=1}^3 G_i^2(z_i)} \\ 0,97x_1 + 4,20x_2 + 1,30x_3 + 3,10x_4 + 2,13x_5 \leq 2710 \\ x_1 + x_2 + x_3 + x_4 + x_5 = 1692 + 170z_1 \quad (4) \\ 7,00x_1 + 12,60x_2 + 1,30x_3 + 7,35x_4 + 5,47x_5 = 7200 + 720z_3 \\ x \geq 0$$

Näherungswerte für die optimalen Lösungen x^* von (4) für verschiedene Werte des Risikoparameters \mathcal{J} und die durchschnittlichen Gesamtkosten $\mu(x^*)$ sind in der folgenden Tabelle angegeben.

In der ersten Zeile (E.P.) steht die optimale Lösung des deterministischen Problems (3). Die folgende Zeile ($\mathcal{J}=0$) enthält die optimale Lösung des Kompensationsproblems:

\mathcal{J}	x_1^*	x_2^*	x_3^*	x_4^*	x_5^*	$\mu(x^*)$
E.P.	328,9	0,0	613,5	0,0	749,6	1659,12
0	29,4	0,0	488,5	0,0	960,8	1682,87
2,0	197,0	87,5	512,6	0,0	697,3	1727,39
3,0	228,6	120,3	508,0	0,0	620,9	1744,39
4,0	245,0	139,9	499,7	0,0	579,7	1756,11

Die Komponenten $\hat{x}_1(R)$, $\hat{x}_2(R)$ bis $\hat{x}_5(R)$ der optimalen Lösung des zu (4) gehörenden Hilfsproblems H_R sind in Abb. 5 bis 9 aufgetragen. Abb.10 zeigt die monoton wachsenden Funktionen $\alpha(\hat{x}(R))$ und die monoton fallenden Funktionen $\mu(\hat{x}(R))$ für verschiedene Werte von \mathcal{J} .

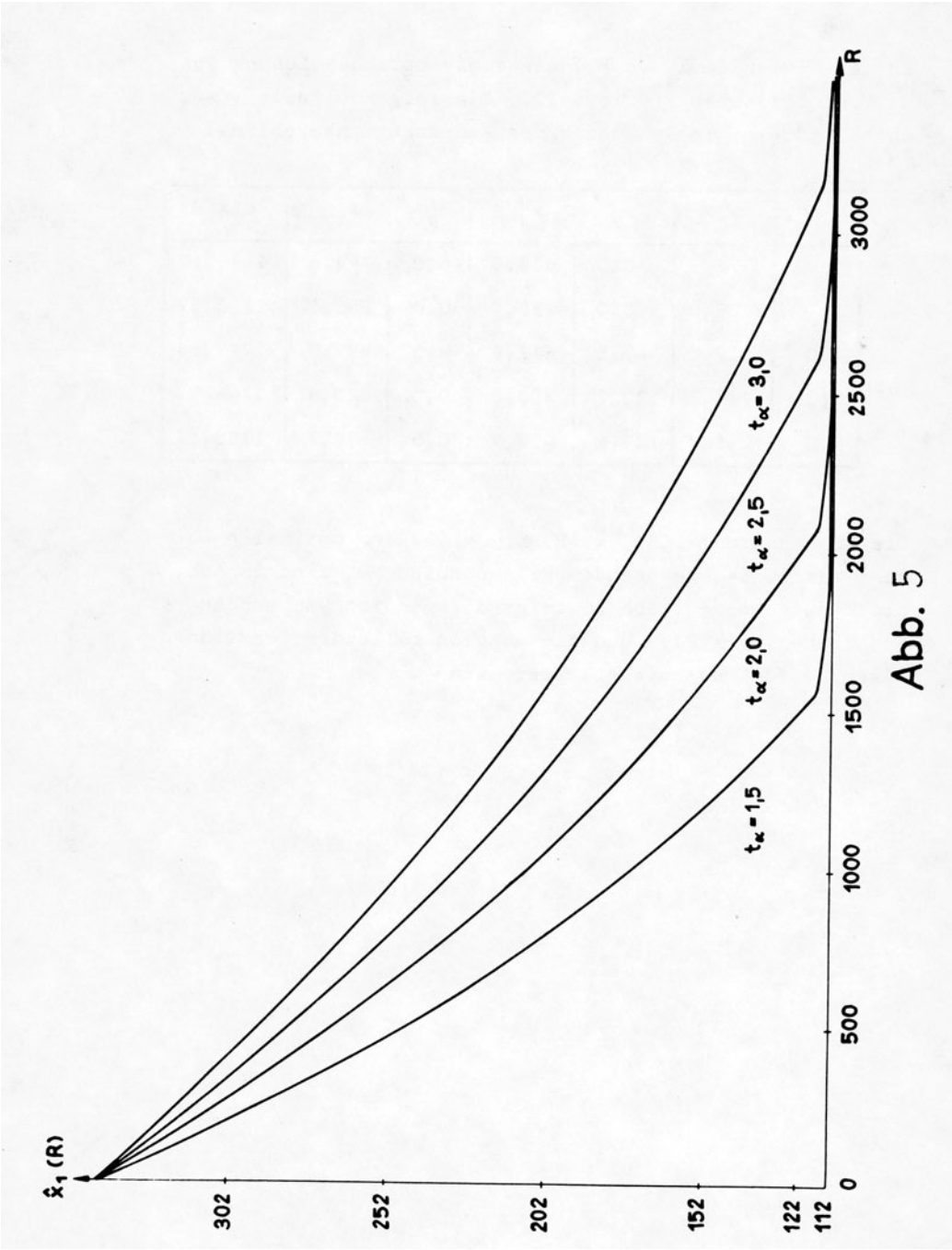


Abb. 5

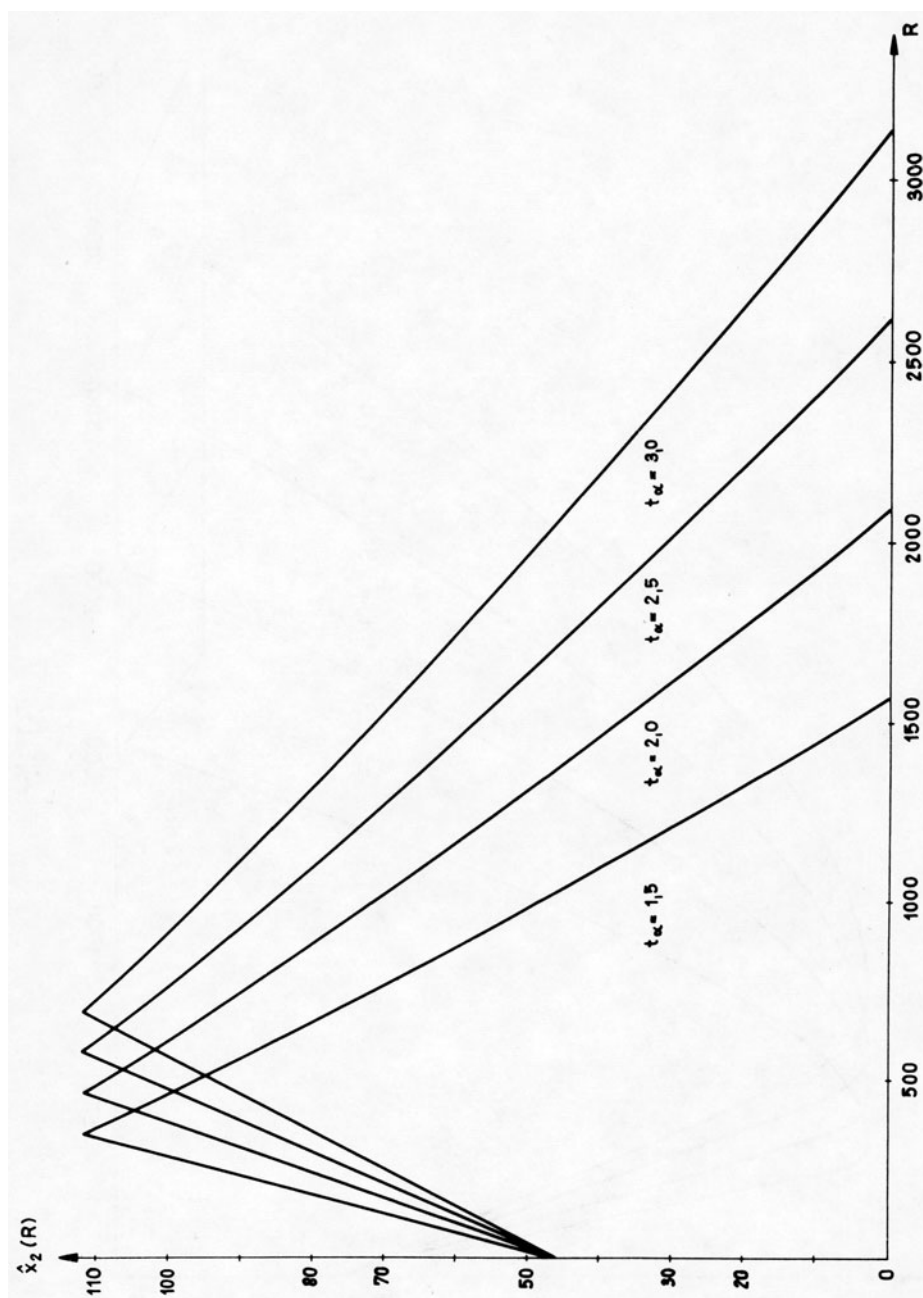


Abb. 6

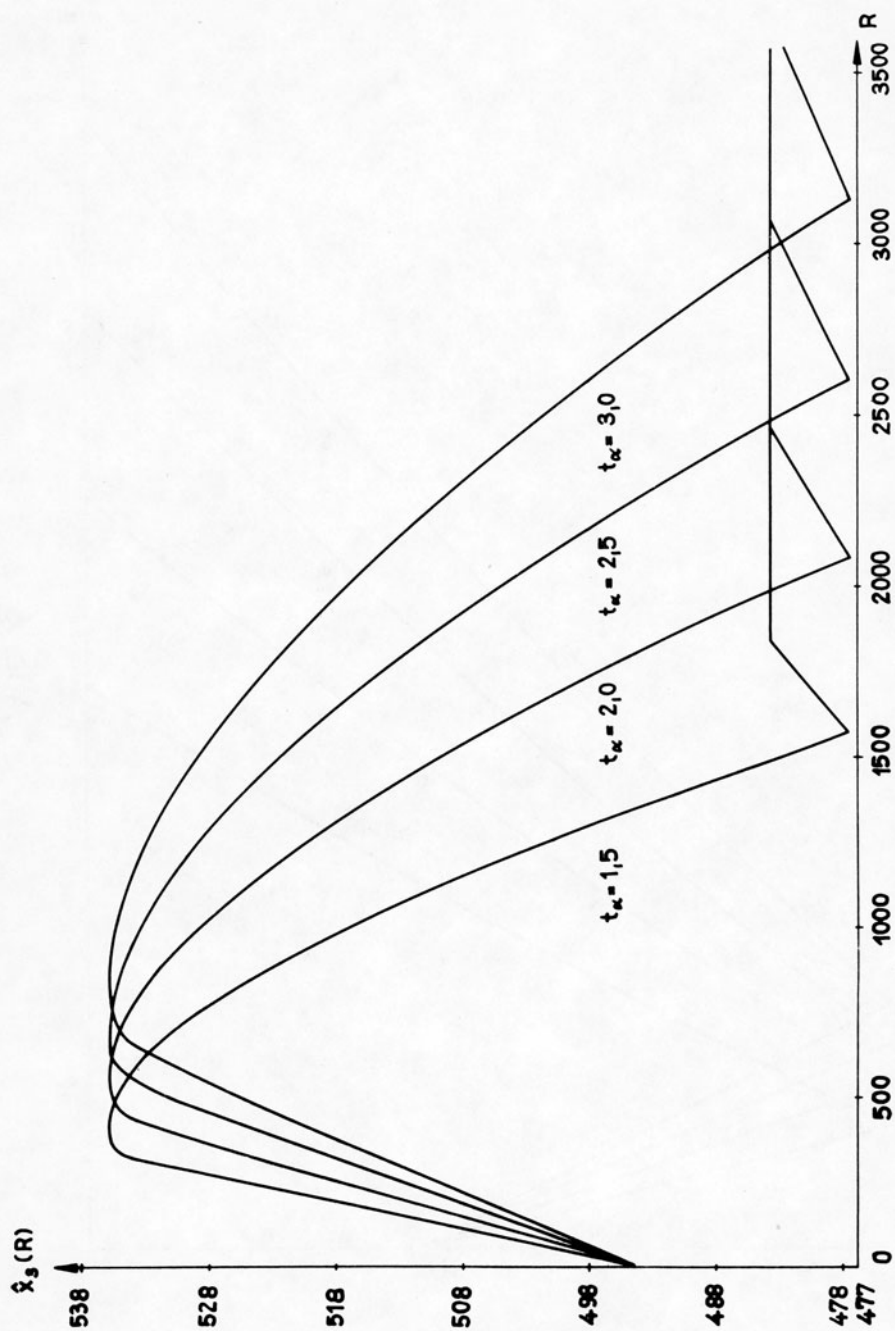


Abb. 7

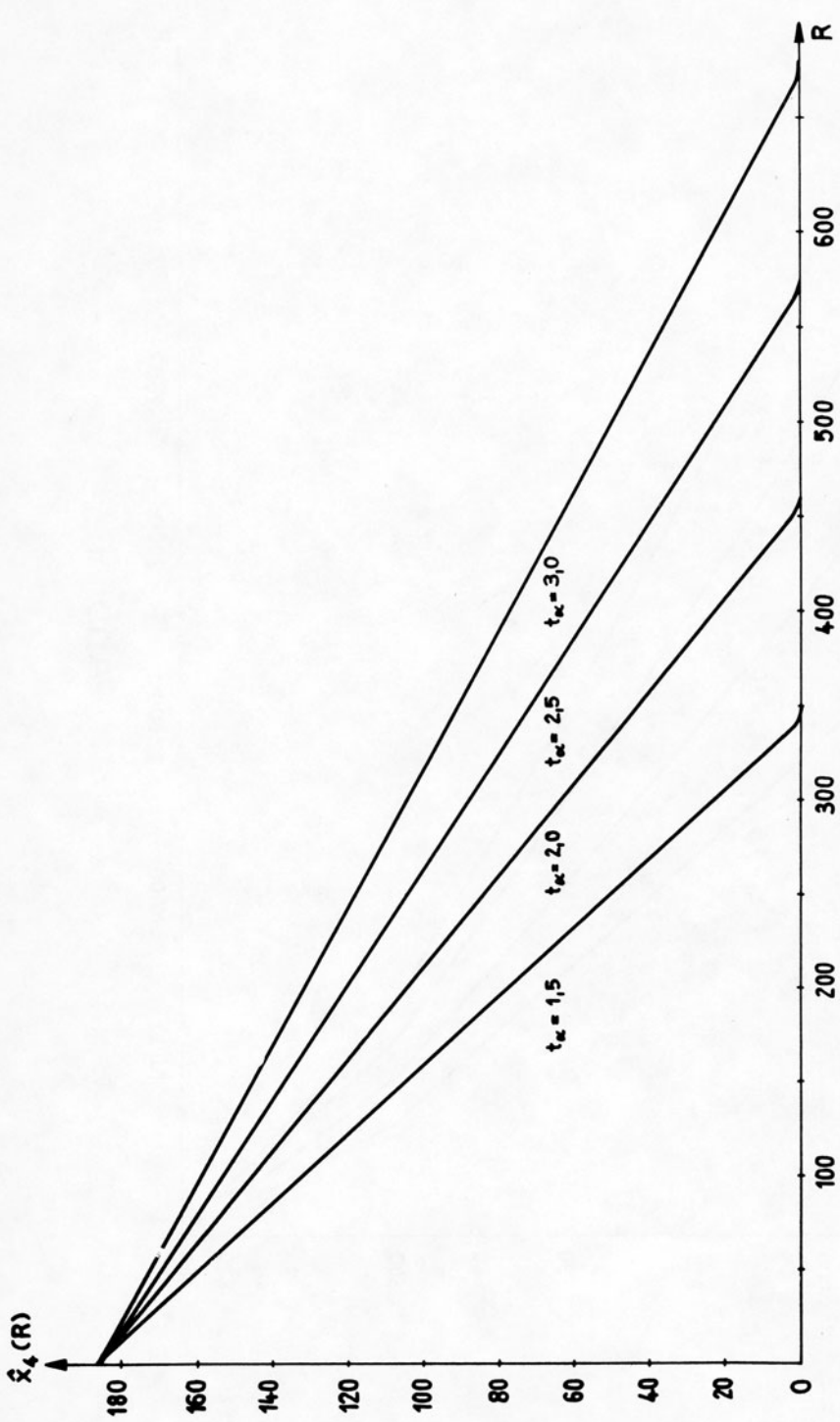
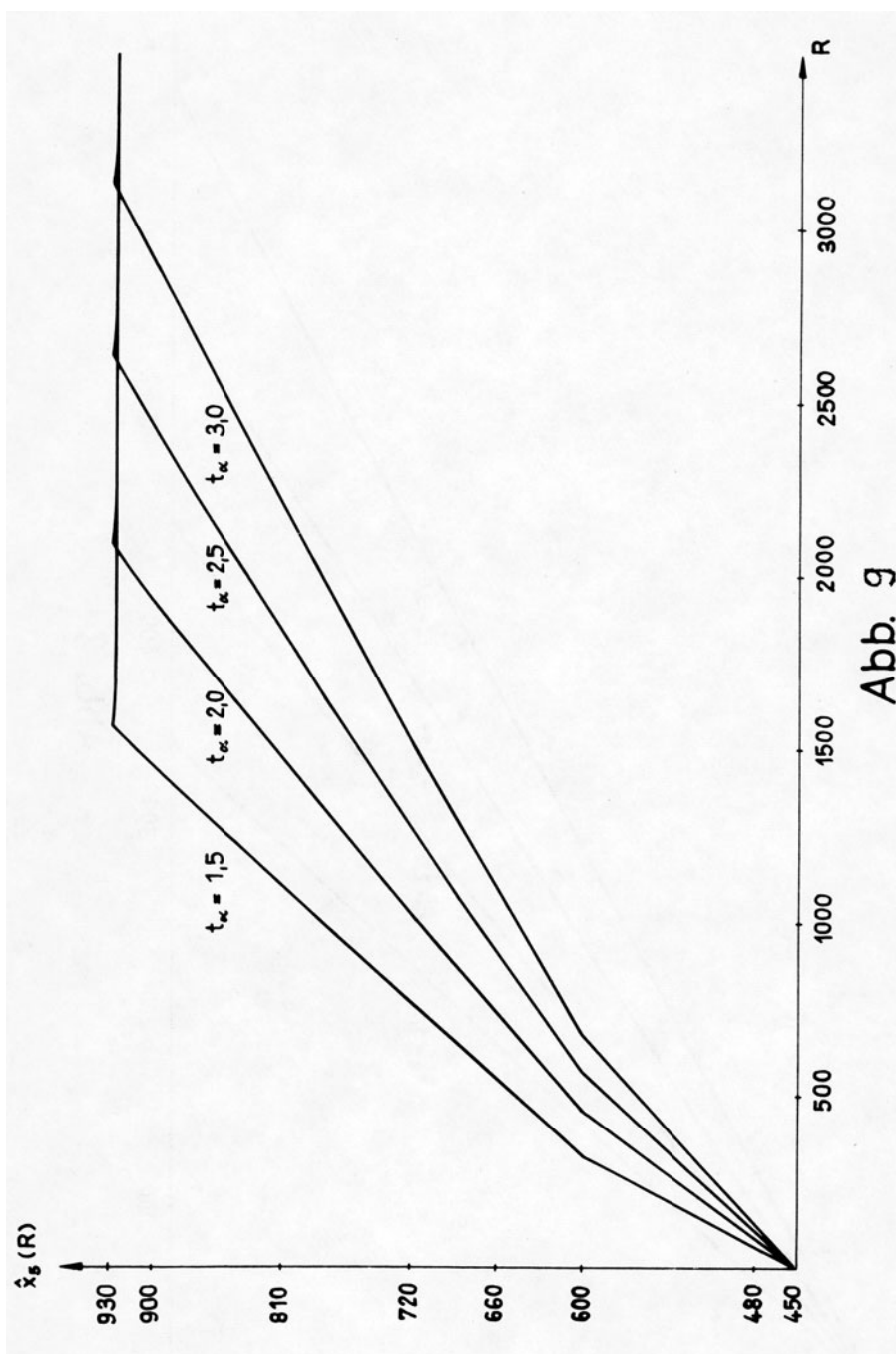


Abb. 8



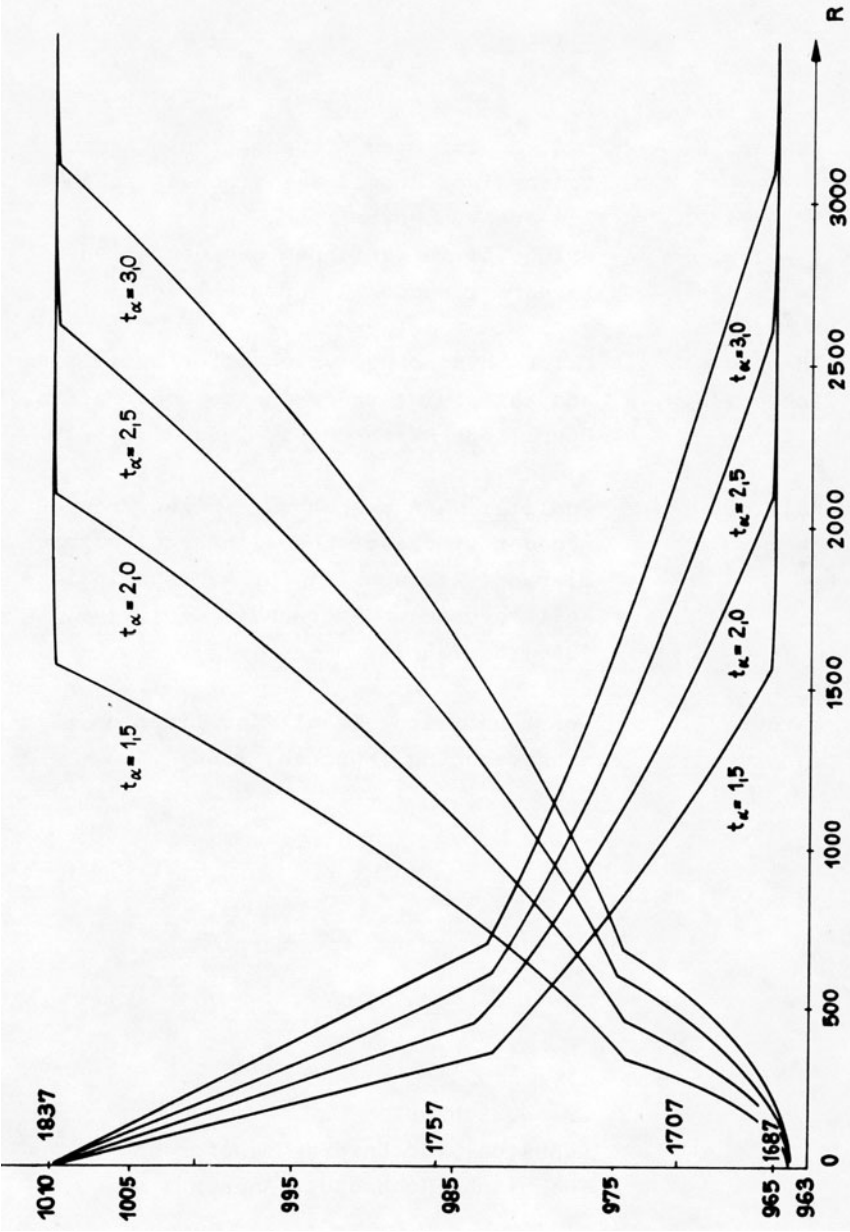


Abb. 10

Literatur

- [B] Bühler, W. Ein kombiniertes Kompensations-Chance-Constrained Modell der stochastischen linearen Programmierung erscheint in Verfahren des Operations Research, Band XII, Hrsg. R. Henn
- [C] Charnes, A. u. Cooper, W.W. Deterministic equivalents for optimizing and satisfying under chance constraints, Operations Research 11, 1963, S. 18-39
- [K] Kall, P. Qualitative Aussagen zu einigen Problemen der stochastischen linearen Programmierung, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 6, 1966, S. 246-272
- [M] Marcus, P. Der ökonomische Inhalt der Linearen Planungsrechnung, München, 1966

Dr. W. Bühler
Lehrstuhl für Unternehmensforschung
Technische Hochschule Aachen

Betriebliche Probleme

Ein Simulationsmodell zur integrativen Unternehmensplanung

von H. Meyhak, Mannheim

Ein Simulationsmodell zur integrativen Unternehmensplanung

I. Zur Problemstellung

A. Corporate Modeling - Unternehmensmodelle

Die Unternehmensplanung beinhaltet das systematische Studium der Auswirkungen verschiedener Vorgehensweisen auf das Unternehmen und dessen Umwelt. Sie unterstützt mit der Analyse des Unternehmensverhaltens bei verschiedenen Handlungsweisen die Entscheidungsfindung im Unternehmen. Gesamtplanungsmodelle dienen dazu, die Auswirkungen möglicher Handlungen auf das Unternehmen quantitativ darzustellen und sollen schließlich helfen, diejenigen Vorgehensweisen auszuwählen, die das Unternehmensziel am besten erfüllen.

Die Betriebswirtschaftslehre bemüht sich seit langem um Modelle, die das Unternehmen als Ganzes abbilden, um die Auswirkungen verschiedenartiger Entscheidungen in ihrer Interdependenz auf das Gesamtunternehmen durchrechnen zu können.

In den letzten fünf Jahren ist, wohl bedingt durch

1. die mehr quantitativ orientierte und mathematische Ausbildung der in die Stabsabteilungen und Führungspositionen der Unternehmen nachrückenden Kräfte,
2. den zunehmenden Einsatz von leistungsfähigen Computern mit sowohl großen externen als auch großen internen Speichern,
3. den härteren Wettbewerb und die dadurch steigende Komplexität der Entscheidungen

auch von der Wirtschaftspraxis der Ruf nach Gesamtplanungsmodellen laut geworden. Das Management hat erkannt, daß eine formale Planung notwendig ist, daß sich nur mit Computermodellen eine laufende, schnelle Revision der Pläne in Anpassung an die veränderten Umweltbedingungen wirtschaftlich durchführen läßt und daß beim Fehlen von exakten Ausgangsdaten für die Planung nur mit Simulationsmodellen eine adäquate Planung möglich ist. Die Planer im Unternehmen sehen die Vorteile des Einsatzes von Computermodellen für die Prüfung ihrer Vorschläge.

Modelle der Unternehmensforschung haben in die Unternehmen Eingang gefunden. In der Regel mußten aber allgemein, entweder um des mathematischen Ansatzes willen oder wegen der beschränkten Computerkapazität, Simplifizierungen der Problemstellung vorgenommen werden oder die Planungsmodelle auf Teilbereiche des Unternehmens beschränkt bleiben.

Die Verbilligung der Kernspeicherkapazität und nicht zuletzt auch die Diskussion bei dem Aufbau von Modellbanken für Management-Informationssysteme haben die Arbeit auf dem Gebiet der Gesamtplanungsmodelle forciert.

B. Zum Stand der Entwicklung von Gesamtplanungsmodellen

In den letzten zwei Jahren sind eine Reihe von Modellen zur Gesamtplanung im Unternehmen in der Literatur beschrieben worden. Große Unternehmen wie Xerox, Sun Oil, Anheuser-Busch, IBM, New York Life Insurance Company u.a. haben Unternehmens-

modelle konzipiert¹⁾. Nach außen dringen aber keine detaillierten Informationen über die Struktur, den Anwendungsbereich der Modelle, die eingehenden Daten, die Annahmen über das Unternehmensverhalten und über die Ausgabedaten²⁾.

Die m.E. wohl beste und umfassendste Zusammenstellung von Arbeiten auf dem Gebiete des Corporate Modeling findet sich in den von Albert N. Schrieber herausgegebenen Proceedings der Conference über "Corporate Simulation Models", die im März 1970 von dem College on Simulation and Gaming of the Institute of

Management Science, Providence, Rhode Island und der Graduate School of Business Administration of the University of Washington, Seattle, Washington, veranstaltet wurde. Diese Proceedings enthalten 25 Referate über den Entwurf und die Anwendung von Modellen zur Simulation des Unternehmens als Ganzes. - Der Ausdruck Simulation wird hier im Sinne von "Planung" gebraucht. Es werden sowohl Modelle der deterministischen Simulation als auch solche der mathematischen Programmierung beschrieben.

Auf dieser Konferenz gab Gershefski einen ausführlichen Bericht über den Stand der Entwicklung auf dem Gebiete des Corporate Modeling³⁾. Er basiert auf der Auswertung eines Fragebogens, der Anfang des Jahres 1969 an die dem Planning Executive Institute angeschlossenen Unternehmen versandt worden war. Die Umfrage zeigt u.a.⁴⁾:

-
- 1) Naylor, Thomas H.: Corporate Simulation Models and the Economic Theory of the Firm, in: Schrieber, Albert N.:(ed.): Corporate Simulation Models, Seattle 1970, S. 1-25, hier S. 19.
 - 2) Vgl. ebenda S. 19.
 - 3) Gershefski, George W.: Corporate Models - The State of the Art, in: Schrieber, a.a.O., S. 26-42.
 - 4) Vgl. ebenda S. 41.

- 95 % der bekannten und in der Entwicklung befindlichen Modelle lag als Struktur ein Simulationsmodell vom Fall-Typ zugrunde,
- 5 % waren Modelle der Mathematischen Programmierung.
- 80 % der Modelle waren deterministischer Art,
- 65 % der Programme waren in FORTRAN,
- 20 % in COBOL,
- 4 % in PL1 und 2 % in DYNAMO geschrieben⁵⁾.
- 65 % der Entwicklungen begannen mit aggregierten Modellen,
- 35 % gingen vom Teil zum Ganzen vor, d.h. Funktion nach Funktion wurde in das Modell einbezogen⁶⁾.

Nach Gershefski betrug der Arbeitsaufwand für die Konzeption und Konstruktion der Modelle zwischen 0,5 und 23 Mann-Jahren, im Mittel 3,5 Mann-Jahre⁷⁾. Davon entfielen auf die allgemeine Konzeptionsphase 25 %, auf die Datensammlung und Datenanalyse 25 %, auf die Entwicklung eines Computermodells 40 % und auf die Implementierung 10 % des Zeitaufwandes⁸⁾.

5) Vgl. ebenda S. 42.

6) Vgl. ebenda S. 41.

7) Gershefski, George W.: Corporate Models - The State of the Art, in: Schrieber, a.a.O., S. 39.

8) Vgl. ebenda S. 40.

Heute existieren erst wenige Modelle, bei denen eine Interaktion zwischen dem Planer und dem Modell auf On Line Basis stattfinden kann. Für die Zukunft wird allerdings ein höherer Anteil dieser Art von Modellen erwartet. Der Planer gibt dann die Daten des Zustandes des Systems Unternehmen am Planungszeitpunkt und die Entscheidungsparameter in das Modell ein, analysiert die für das Systemverhalten kritischen Parameter und verändert diese 'kritischen Parameter solange bis das Simulationsergebnis bzw. das Systemverhalten mit den von den Unternehmen angestrebten Zielen verträglich ist. Heute sind die auf On Line Basis benutzten Modelle in der Regel als Alternativen-Tester konzipiert. In der Zukunft sind Modelle zu erwarten, in deren Ablaufalgorithmen Adaptions- und Optimierungsroutinen eingebaut sind.

Andersen⁹⁾ glaubt, daß infolge verbesserter Optimierungsalgorithmen der Anteil der Modelle der Mathematischen Programmierung zunehmen wird. Er weist insbesondere auf den Einsatz der allgemeinen upper bounding technic hin, mit der es möglich ist, bei bestimmten Strukturen den Rechenaufwand schon heute entscheidend zu reduzieren.

9) Anderson, Richard L.: Building Corporate Planning Models Using Mathematical Programming Techniques - A Case Study, Paper presented to 11th American Meeting of The Institute of Management Sciences, Los Angeles, California, Oct. 20, 1970.

Corporate Models werden im Rahmen von Management-Informationssystemen (MIS) eingesetzt. Bei den MIS werden in Anlehnung an Mason¹⁰⁾ unterschieden (Bild 1):

1. Datenbank-Informationssysteme,
2. Voraussagende Informationssysteme,
3. Optimale Entscheidungen errechnende (vorschlagende) Informationssysteme,
4. Entscheidungen errechnende und Handlungen anordnende Informationssysteme (Decision taking Information System).

Das voraussagende Informationssystem stellt eine Erweiterung des reinen Datenbank-Informationssystems um eine Modell- oder Methodenbank mit einem System von Rechenmodellen und/oder Rechenalgorithmen dar. Für eine Reihe von Annahmen über unternehmensexterne und über unternehmensinterne Daten (Parameter) und über mögliche alternative Handlungen werden mit Hilfe der Modelle der voraussichtliche Zustand und/oder die Verhaltensweise des Unternehmenssystems errechnet. Die Bewertung der Auswirkungen der in das Modell eingegebenen Entscheidungsparameter im Hinblick auf das System Unternehmen, wie das Erreichen eines gewünschten Ziel- oder Anspruchsniveaus oder das Systemverhalten erfolgt nicht durch den Modellalgorithmus, sondern muß von den Entscheidungsträgern vorgenommen werden.

Beim optimalen Entscheidungen errechnenden Informationssystem sind in das Modell Optimierungsalgorithmen eingebaut. Dem Verantwortlichen bleibt es dann überlassen, ob er die als optimal

10) Mason, Jr., Richard O.: Basic Concepts for Designing Management Information Systems, Graduate School of Business Administration, University of California, Los Angeles, AIS Research Paper No. 8, Oct. 1969.

errechneten Alternativen seiner Entscheidung zugrundelegt.

Das von mir konzipierte Modell trägt im wesentlichen die Merkmale des Typs: Voraussagendes Informations-System.

II. Der strukturelle Aufbau des Modells

A. Die spezifische Problemstellung des Modells

Das Modell ist ein deterministisches Simulationsmodell für ein Unternehmen mit einem breiten Produktionsprogramm, mit Werkstattfertigung und anschließender Montage der Baugruppen.

Das Unternehmen fertigt im Wochenzyklus vier verschiedene Erzeugnisarten, die aus mehreren Bauelementen zusammengesetzt werden. Den strukturellen Aufbau des Erzeugnisses 100 zeigt das Bild 2. Die für verschiedene Auslieferungstermine zu fertigenden Enderzeugnisse und auch deren Bauelemente sind nicht gegeneinander austauschbar. Eine verspätete Auslieferung reduziert beträchtlich und nachhaltig sowohl das gegenwärtige als auch das zukünftige Absatzpotential. Andererseits darf mit der Bearbeitung der Bauelemente im allgemeinen nur eine bestimmte, relativ kurze Zeit vor dem fixierten Auslieferungstermin der Enderzeugnisse begonnen werden. Eine eventuelle Lagerung der Erzeugnisse erfolgt nur bis zum vorgesehenen Auslieferungstermin. Die nicht abgesetzte Menge ist wertlos und wird vernichtet.

Zusätzliche zyklisch oder auch nur einmalig zu fertigende Erzeugnisarten können in "beliebiger" Anzahl, d.h. soweit in der vorliegenden Programmversion die Kapazität der im Kernspeicher stehenden Bedarfsplan-Matrix mit den in Fertigungslose aufgelösten Bedarf an Erzeugnissen reicht, als zusätzlicher Bedarf verarbeitet werden. Für das Modellunternehmen wurde diese Art des Produktionssortimentes gewählt, um für den Prototyp des Simulators den Arbeitsaufwand für die Entwicklung einer Bestands-

und Bestellrechnung von mehrfach verwendbaren Bauelementen einzusparen.

Das Unternehmen verfügt über

- 8 verschiedene Arten von Betriebsmitteln,
- 6 verschiedene Arten von Arbeitskräften und verwendet
- 5 verschiedene Arten von Material.

Das Simulationsmodell wurde für ein voraussagendes Informations-System erstellt, es gibt also als Alternativen-Tester auf "Was geschieht, Wenn"-Fragen eine Antwort. Später sollen dann, in einer verbesserten und erweiterten Version des Simulators,

1. Adaptionenalgorithmien und
2. Optimierungsalgorithmen für Teilbereiche des Unternehmens
in das Modell eingebaut werden.

Der Simulator soll weiter in der vorliegenden Version - als Prototyp - zum Erarbeiten der wesentlichen Strukturen eines Gesamtplanungsmodelles des Unternehmens und zum Testen von Adaptionen- und Optimierungsalgorithmen dienen. Das erforderte einerseits, das Modell so stark zu detaillieren, daß die wesentlichen Aufbau- und Ablaufstrukturen erkennbar wurden, andererseits aber, da es sich um ein Testmodell handelt, den Arbeitsaufwand möglichst gering zu halten. Deshalb wurde für den Prototyp eine reine Kernspeicherversion gewählt. In der endgültigen, verbesserten Version des Simulators sollen die Systemzustands-Matrizen auf Platten gespeichert werden.

Weiter war bei der Konzeption des Modells zu berücksichtigen, daß es auch als Management Game mit nur geringen Änderungen eingesetzt werden kann. Es müssen also Eingriffe in den Systemablauf auch an außerhalb des Planungsnullpunktes gelegenen Zeitpunkten möglich sein.

B. Zum Vorgehen beim Aufbau des Modells

Zwei Arten des Vorgehens beim Aufbau eines Corporate Models stehen zur Auswahl:

1. der top down approach,
2. der bottom up approach.

Als Beispiel für einen top down approach seien erwähnt die Simulationsmodelle vom Typ "INDUSTRIAL DYNAMICS", heute auch "SYSTEMS DYNAMICS" genannt, mit denen die Verhaltensanalyse kontinuierlicher Systeme im Wege des Vorwärtseintrüttelns nach bestimmten Regeln vorgenommen wird. Modelle dieses Typs haben im allgemeinen einen sehr hohen Abstraktionsgrad.

Der top down approach, d.h. das Arbeiten mit vorwiegend aggregierten Größen, ist zweifellos eine einfache, schnelle, wenig zeitaufwendige und übersichtliche Vorgehensweise. Der Aufbau eines Modells im Wege des bottom up approaches ist zwar arbeitsaufwendiger als beim top down approach. Das Arbeiten mit diesen Modellen ergibt aber wesentlich aussagefähigere und eindeutige Ergebnisse, die ohne zusätzliche Interpretationen verwendbar sind. Die Modelle lassen einen größeren Katalog möglicher Fragestellungen auf allen Stufen der Unternehmenshierarchie zu.

In den vorerwähnten Proceedings vertritt die Mehrzahl der Autoren die Ansicht, daß der Trend zur Entwicklung von Gesamtplanungsmodellen in Richtung des bottom up approaches geht, also zu Modellen detaillierterer Art. Für die vorliegende Arbeit wurde deshalb ein bottom up approach gewählt.

Für die Arbeit stand eine Siemens 4004/45 mit zunächst 128 K Bytes, später mit 256 K Bytes zur Verfügung. Es hätten eine aus FORTRAN Subroutinen bestehende DYNAMO Version, GASP, eine ebenfalls aus FORTRAN Subroutinen bestehende SIMSCRIPT ähnliche Simulationssprache, SIMULATE und PASS verwendet werden können.

Die Durchsicht der Literatur, insbesondere des Buches von Schrieber, ergab, daß FORTRAN in der Mehrzahl der Fälle bei der Entwicklung von Corporate Models verwendet worden war. Um ein effizientes Programm zu schreiben, mit dem beschränkten Speicherraum beim Prototyp auszukommen und um die Strukturen der Rechnung besser erkennen zu können, wurde FORTRAN als Programmiersprache gewählt.

C. Das dem Modell zugrundeliegende Systemkonzept

Um die Konstruktion des Modells schnell in den Griff zu bekommen, um mit einem zeitlich verträglichen Aufwand den Prototyp entwickeln zu können, wurde zunächst die Struktur des Unternehmens und des Planungsablaufs analysiert und dann eine einheitliche Struktur dem Modellaufbau zugrundegelegt.

Das System Unternehmen kann als ein System von Potentialen aufgefaßt werden, die im Zeitablauf in ihrer Größe Änderungen unterliegen. Potentiale sind alle diejenigen Größen im Unternehmen (wie die Produktionsfaktoren, statistische Aufzeichnungen und allgemeine Kennziffern = Informationsbestände), die bei einem Stillstand der Zeit erhalten bleiben.

Dazu gehören u.a.:

der Materialbestellbestand,

der effektive Materialbestand,

der Bestand an Debitoren und

die statistischen Angaben über die Faktorauslastung, den Trend

der Faktorauslastung, den Trend im Auftragseingⁿag,

die Saisonfaktoren,

das Akquisitorische Potential usw.

Ein Potential einer bestimmten Art besteht aus einer beliebigen Anzahl von Elementen, die in mindestens einer, die Zugehörigkeit

zum Potential kennzeichnenden Eigenschaft gleich sind, im übrigen aber eine beliebige Anzahl verschiedener permanenter und temporärer Eigenschaften aufweisen können. So kann sich z.B. das Potential Materialbestellbestand einer bestimmten Materialart aus mehreren Bestellungen von verschiedenem Umfang, unterschiedlichem Einzelpreis und Lieferterminen zusammensetzen. Das Potential effektiver Bestand an Betriebsmitteln einer bestimmten Art kann aus zu verschiedenen Zeiten und zu verschiedenen Preisen angeschafften Betriebsmitteln mit einem unterschiedlichen Buchwert und unterschiedlicher Restnutzungsdauer bestehen. Die Betriebsmittel können sich darüber hinaus in der möglichen täglichen Nutzungsdauer, den marginalen Nutzungskosten unterscheiden, durch Aufträge belegt oder nicht belegt sein.

Die Elemente der Potentiale werden in dem System Unternehmen als in einer Warteschlange stehend aufgefaßt. Das Potential kann nach Definition im Extremfall auch aus nur einem Element bestehen, wie z.B. das Potential Kassenbestand. Von den den Elementen zugeordneten Eigenschaften hängt es ab, ob und wann die Elemente die Warteschlange verlassen, welchem Potential im Unternehmen sie anschließend zugeordnet werden, oder ob sie untergehen. Die Anzahl der Elemente eines Potentials ist im Modell nur ganzzahlig diskret veränderbar. So stellen sowohl die vor den Betriebsmitteln auf Bearbeitung wartenden Aufträge als auch die Betriebsmittel selbst ein Potential und damit eine Warteschlange dar. Die letzteren warten auf den Abgang aus dem Unternehmen.

Die Warteschlange ist rein gedanklich zu sehen. So bleiben die Aufträge vor den Betriebsmitteln, unabhängig von dem Wechsel ihrer Priorität im Hinblick auf die Bearbeitung auf der Maschine, an dem gleichen Platz in der Matrix (Speicherstelle) stehen. Die Bewegung der Warteschlange erfolgt allein gedanklich durch

die Änderung der Anzahl der Elemente der Warteschlange beim Zugang von Aufträgen bzw. beim Abgang infolge der Bearbeitung auf der Maschine und durch Änderung der Priorität der Aufträge.

Die deterministischen und stochastischen Simulationen basieren allgemein auf einem Vorwärtstritteln nach bestimmten, vorgegebenen Regeln. Der Programmablauf erfolgt dabei jeweils in dem Zyklus:

1. Erfassen der Informationen über den Zustand der Elemente eines Potentials (Warteschlange),
2. Verarbeiten der Informationen über den Zustand der Potential-elemente,
3. Vergleichen der verarbeiteten Informationen über die Elemente mit dem angestrebten Zustand der Elemente, im weiteren Sinne mit der Zielvorstellung, bzw. der Voraussetzung für eine Aktivität und Ermitteln der Art der Annäherung an den gewünschten Zustand,
4. Eventuell Durchführen der Handlung an den Potential-Elementen,
5. Fortfahren mit 1.

Die Programmschritte sind nichts anderes als die Strukturelemente einer Rückkopplungsschleife.

Forrester stellt in seiner INDUSTRIAL DYNAMICS Konzeption das Vorhandensein von Feedback Loops besonders heraus; ja, er legt seinem System als Strukturelemente ausschließlich Feedback Loops zugrunde.

In den anderen Simulationskonzeptionen, wie GPSS, GASP, SIMSGRIPT usw. werden diese Feedback Loops dagegen nicht expressis verbis hervorgehoben.

Die aus physischen und Informations-Beständen bestehenden Potentiale-Warteschlangen des Unternehmens werden in einem Warteschlangen-Netzwerk zusammengefaßt, dem ein Informations-

netzwerk überlagert wird. Mit Hilfe des Informationsnetzwerkes werden die Informationen über die Elemente der Potentialgrößen (Warteschlangen) einzeln oder in aggregierter Form erfaßt, z.T. verarbeitet mit Hilfe von Multiplikatoren, Funktionen usw., und schließlich die aus den verarbeiteten Informationen abgeleiteten Entscheidungen in Handlungen transformiert, die zu einer Änderung der Potentiale führen. Eine Bewegung der Elemente des Potentials braucht dabei nicht zu erfolgen.

Die Änderungen der Potentiale erfolgen in einem von der Art des Potentials abhängigen zeitlichen Rhythmus. So werden z.B. Materialbedarfsrechnungen zur Ermittlung des Bestellumfangs wöchentlich vorgenommen, Materialabgänge an die Produktion erfolgen einmal täglich (Abtastregelung nach DIN 19226, Ziffer 4.3). Diese Ausführungen zeigen, daß dem Simulator des Systems Unternehmen ein System kybernetischer (geregelter) Warteschlangen im Modell zugrundeliegt.

D. Die Moduln des Unternehmens-Simulators

Das Programm des Unternehmens-SIMulators 02 (USIM02) besteht aus drei Teilen:

1. dem Ladeprogramm des Systems,
2. dem Simulationsprogramm und
3. dem Reportgenerator.

1. Das Ladeprogramm

Vor dem Simulationslauf ist das System zu laden mit den Daten, die den Zustand der Potentiale des Unternehmens (Warteschlangen) am Planungsnullpunkt beschreiben. Dazu gehören Angaben über:

den Materialbestellbestand (Art, Bestellmenge, Lieferdatum
und Preis),

den effektiven Materialbestand (Art, Menge und Bilanzbewertung),

den Betriebsmittelbestellbestand	(Art, Menge, Anschaffungspreis, Lieferdatum),
den Betriebsmittelbestand	(Art, Menge, Anschaffungspreis, Nutzungsdauer, Restnutzungsdauer, tägliche Arbeitszeit),
den Arbeitskräftebestand	(Art, Menge, Kosten pro Stunde, tägliche Arbeitszeit),
das Akquisitorische Potential	(Umfang, Restnutzungsdauer, Kosten pro Einheit),
den Kreditorenbestand	(Umfang, Fälligkeit, Kosten),

den effektiven Bedarf und den geschätzten Bedarf an Erzeugnissen in der vor dem Planungsnullpunkt gelegenen Periode, die Anpassungskonstanten, die Saison- und Trendkomponenten für die Bedarfsschätzung mit Hilfe der Methode des Exponential Smoothing und die Arbeitsplan-Matrix für die zyklisch zu fertigenden Erzeugnisse.

2. Das Simulationsprogramm

Das "eigentliche" Simulationsmodell (Bild 3) besteht aus einer Reihe von Modulen, die in einem unterschiedlichen zeitlichen Rhythmus - Stunden, Tage, Wochen oder Monate - aufgerufen werden. Den Kern der Informationen im Bereich der Produktionsplanung enthält die Bedarfsplan-Matrix. Wöchentlich werden durch Aufruf der Subroutine BEDRF1 der zyklische Erzeugnisbedarf geschätzt und in die Bedarfsplan-Matrix geschrieben. Sodann erfolgt eine Auslösungsrechnung der Erzeugnisse. Die dazu notwendigen Daten, wie die Anzahl der bis zur Fertigstellung des Erzeugnisses noch notwendigen Arbeitsgänge (Arbeitsgang-Nr.), die Stückzeit, die Rüstzeit, die zur Bearbeitung notwendige Betriebsmittel- und Arbeitskraftart und die Art und Anzahl der Auflösungspeditionen werden der Arbeitsplan-Matrix (Bild 4) entnommen.

Die Bedarfsplan-Matrix (Bild 5) enthält u.a. folgende Angaben:

die Auftragsnummer, bezogen auf das Erzeugnis (Spalte 1),
 den Auslieferungstermin des Erzeugnisses in Tagen (Spalte 2),
 die Erzeugnisnummer (Spalte 3),
 die Bestellnummer (Spalte 4),
 die Anzahl der zu fertigenden Elemente (Spalte 5),
 die Rüstzeit (Spalte 6),
 die Stückzeit (Spalte 7),
 die Betriebsmittelart (Spalte 8),
 die Arbeitskräfteart (Spalte 9),
 den spätest möglichen Beginn der Bearbeitung des Auftrags (Sp.12),
 den spätest möglichen Zeitpunkt der Fertigstellung des Auftrags
 (Spalte 13),
 die Art und die Anzahl der in das Element eingehenden Elemente
 bzw. des eingehenden Materials (Spalte 15, 16, 18, 19, 21, 22,
 24, 25),
 die Anzahl der bis zur Fertigstellung des Erzeugnisses noch notwendigen Arbeitsgänge (Spalte 27).

Die Kennziffer in Spalte 29 gibt in den beiden letzten Stellen die Spalte, in den vorangehenden die Zeile in der Bedarfsplan-Matrix an, in die das Bauelement bzw. der Auftrag eingeht. In Spalte 28 sind die Kennziffern über den Zustand des Auftrages angeführt. Die Zusammenstellung in Bild 6 enthält die Schlüsselzahlen, die den Zustand des Auftrages kennzeichnen.

Mit Hilfe der Subroutine AUSLS1 werden die Bestände an Betriebsmitteln und Arbeitskräften im Hinblick auf die Auslastung abgetastet und die Statistik des Nutzungsgrades dieser Produktionsfaktoren fortgeschrieben. Weiter wird die Anzahl der vor den Betriebsmitteln auf Bearbeitung wartenden Aufträge und deren Arbeitsumfang erfaßt und gespeichert.

Die Subroutine LOSAB1 prüft in der Betriebsmittelbestands-Matrix (Betriebsmittel-Belegungs-Matrix) (Bild 7) in der Spalte 4 das Bearbeitungsende der Aufträge. Sie gibt dann das Betriebsmittel und die entsprechende Arbeitskraft für eine weitere Belegung frei und vermerkt die Fertigstellung des Auftrages in der Bedarfsplan-Matrix in Spalte 28 durch die Kennziffer "99".

Die Subroutine BETLG1 nimmt die Betriebsmittelbelegung vor. In der Betriebsmittelbestands-Matrix wird nach einem freien Betriebsmittel gesucht, das Vorhandensein einer Warteschlange an Aufträgen vor dieser Betriebsmittelart geprüft und die Priorität der vor den Betriebsmitteln wartenden Aufträge im Hinblick auf die Bearbeitung errechnet. Für den Auftrag mit der höchsten Priorität wird in der Arbeitskräftebestands-Matrix (Bild 8) nach einer freien Arbeitskraft gesucht und die Belegung des Betriebsmittels und der Arbeitskraft vorgenommen. Der Fertigstellungszeitpunkt des Auftrages wird ermittelt, in die Betriebsmittelbelegungs-Matrix geschrieben und die Bearbeitung des Auftrages in der Bedarfsplan-Matrix in Spalte 28 durch die Kennziffer "88" vermerkt.

Die Subroutinen AUSLS1, LOSAB1, BETLG1 können, durch Parameter gesteuert, im ein, zwei, vier oder acht Stunden Rhythmus aufgerufen werden.

Im Anschluß an den Aufruf der Subroutinen AUSLS1, LOSAB1, BETLG1 erfolgt die Abfrage auf das Ende des Tages. Am Tagesende werden die Subroutinen ERZVK1 und WASZU1 aufgerufen. Die Subroutine ERZVK1 sucht in der Bedarfsplan-Matrix nach fertigen End-erzeugnissen und überträgt diese Erzeugnisse in die Erzeugnisbestands-Matrix. Ist der Auslieferungszeitpunkt erreicht, so werden der effektive Bedarf der Erzeugnisse ermittelt, der Bestand an Erzeugnissen gelöscht und die Debitorenbestände fortgeschrieben.

Die Subroutine WASZU1 sucht in der Bedarfsplan-Matrix nach Aufträgen, die zeitlich für die Überschreibung an den Betrieb zugelassen sind, prüft, ob das in diese Aufträge eingehende Material bzw. die eingehenden Bauelemente vorhanden sind, und schreibt die Aufträge dem Bestand der vor den Betriebsmitteln wartenden Aufträge zu. Die entsprechende Kennziffer wird in die Spalte 28 der Bedarfsplan-Matrix gesetzt.

Am Wochenende werden die Material-Bestands- und Bestellrechnung und die Materialdisposition mit Hilfe der Subroutine MATDS1 vorgenommen. Es erfolgen weiter am Wochenende die Finanzdispositionen mit Hilfe der Subroutine FINDS1, die Bedarfsrechnung der Erzeugnisse mit BEDRF1 und die Disposition des Akquisitorischen Potentials (Werbung, Forschung und Entwicklung) mit Hilfe der Subroutine AKQDS1.

Die Disposition des Zu- und Abganges von Arbeitskräften (Subroutine ARBDS1) und von den Betriebsmitteln (Subroutine BETDS1) werden jeweils am Monatsende auf Grund der Bedarfsvorschätzung der Erzeugnisse, der Höhe des Akquisitorischen Potentials, des erwarteten Faktorabganges, des Faktorbestellbestandes, des Faktorbestandes und der Faktorauslastung vorgenommen.

3. Der Reportgenerator.

Sämtliche Potentiale (Warteschlangen) des Unternehmens, also die Betriebsmittelbestands-Matrix mit der Betriebsmittelbelegung die Arbeitskräftebestands-Matrix usw., die Bedarfsplan-Matrix, die Bilanz, die Gewinn- und Verlust-Rechnung und der Finanzplan können, durch Parameter gesteuert, vollständig oder partiell, d.h. die volle Matrix oder in konzentrierter Form, z.B. nur die besetzten Zeilen der Matrix oder aber die Zeilen, die bestimmten Bedingungen genügen, zu jedem beliebigen Zeitpunkt, in jedem beliebigen Rhythmus und an beliebiger Stelle während des Programmablaufes ausgedruckt werden.

Ein vollständiger Ausdruck des Systemzustandes im kürzesten gewählten Abtastrhythmus erfordert einerseits unverhältnismäßig viel Druckerzeit und bringt andererseits relativ wenig Informationen. Um den Stand der Aufträge im Zeitablauf zu verfolgen, genügt es, die Bedarfsplan-Matrix im Tagesrhythmus auszu drucken. Aus ihren Kennziffern lassen sich alle notwendigen Angaben über den Zustand der Aufträge und die Einhaltung der Termine entnehmen. Die übrigen Potentiale und der Status des Unternehmens könnten im wöchentlichen oder monatlichen Rhythmus ausgedruckt werden.

III. Zusammenfassung

Wie am Anfang der Ausführungen bemerkt wurde, handelt es sich bei USIM02 um den Prototyp eines Corporate Models für die Modellbank eines MIS auf der Basis einer reinen Kernspeicherversion, die im wesentlichen dazu diente, mit den strukturellen Gegebenheiten des umfangreichen Problems vertraut zu werden und um bestimmte Routinen eines erweiterten, komplexeren Modells auszutesten. Diesen Zweck erfüllt das Modell.

Nunmehr wird an einem wesentlich umfangreicheren, praxisreifen Modell in der Form einer Plattenversion gearbeitet. Mit diesem Modell sollen kurz- und langfristige Planungsprobleme behandelt werden. Es soll aber auch - um den Grad der Detailliertheit in der Endkonzeption zu verdeutlichen - der kurzfristigen Arbeitsvorbereitung dienen, also Arbeitspapiere für den Betrieb herausdrucken. Der Arbeitsvorbereiter soll an seiner Konsole aus verschiedenen Läufen die zweckmäßigsten Alternativen an Hand geeigneter Kriterien auswählen und zur Grundlage der Vorgabe machen können. Andererseits sollen sich auch strukturierte Entscheidungen des Management mit dem Modell durchspielen lassen.

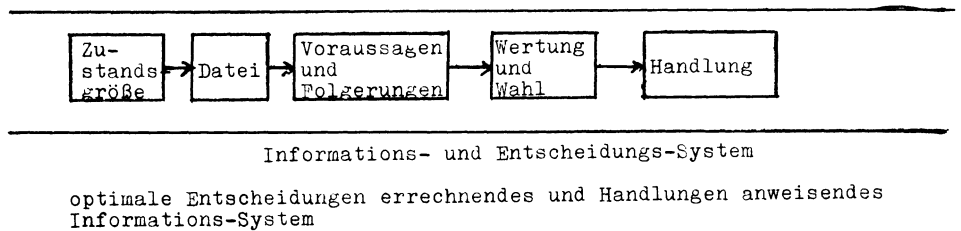
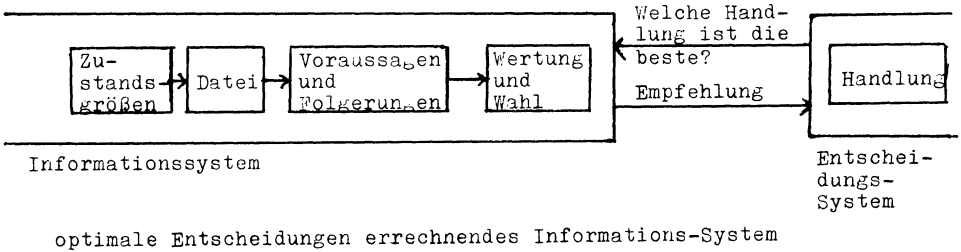
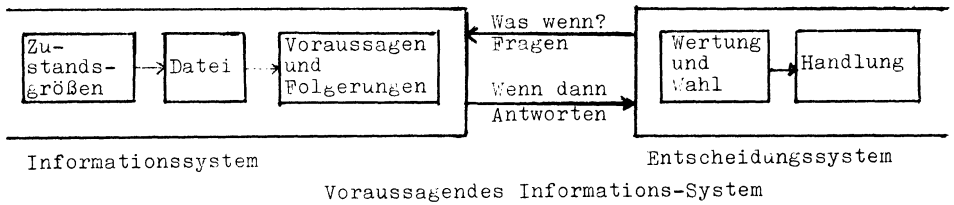
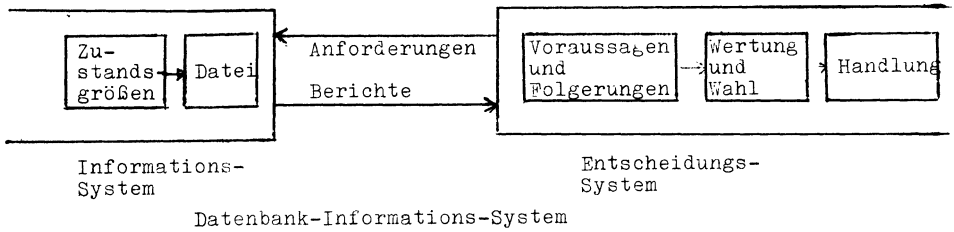


Bild 1

Schematik der Management-Informationssysteme¹⁾

1) entnommen aus: Mason, Jr., Richard O.: Basic Concepts for Designing Management Information Systems, AIS Research Paper No. 8, Graduate School of Business Administration/Division of Research, University of California, Los Angeles, 1969

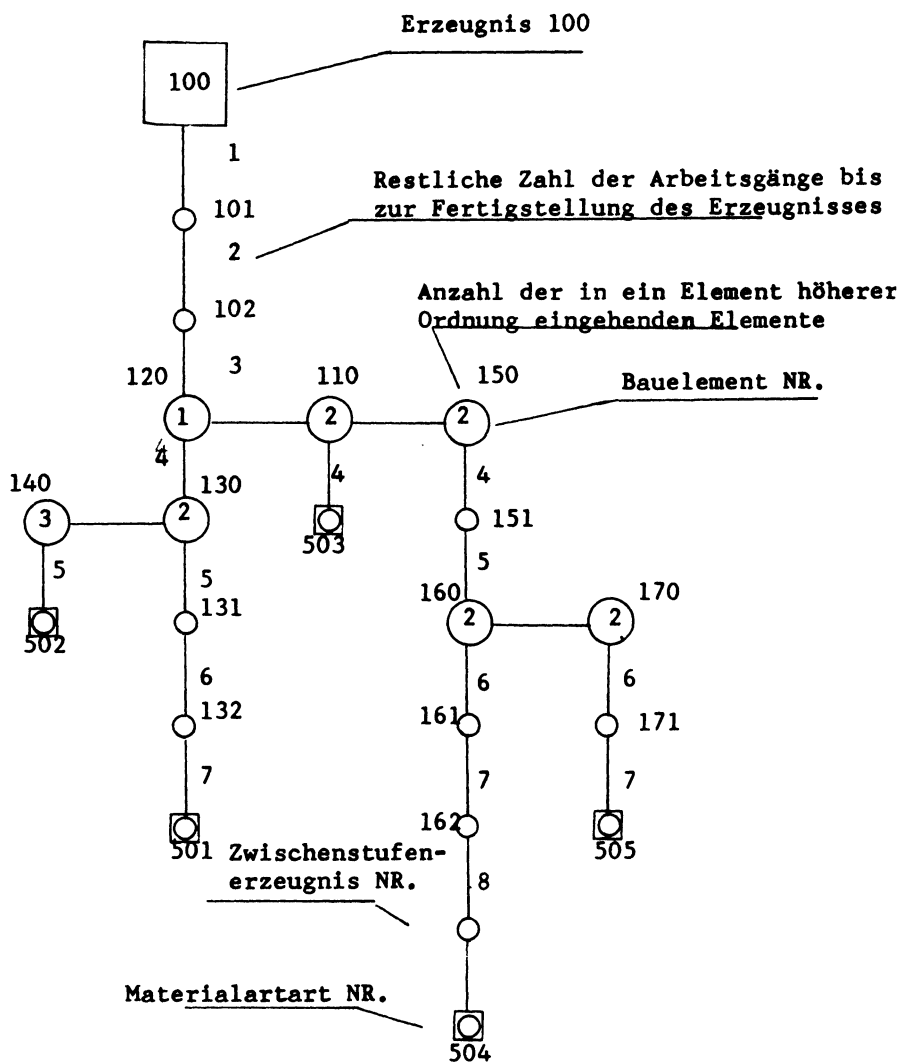


Bild 2: Struktureller Aufbau des Erzeugnisses 100.
 Dreistellige Ziffern mit einer "0" in der 3. Stelle bezeichnen Bauelemente, mit einer "5" an der ersten Stelle Material, die übrigen Zwischenstufenerzeugnisse.

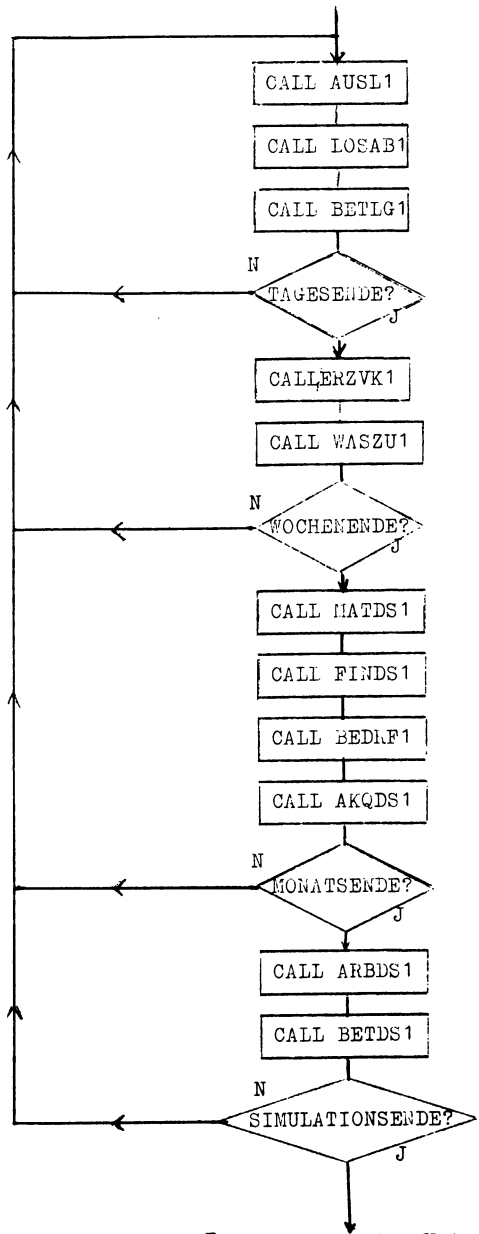


Bild 3: Ausschnitt aus dem Flußdiagramm des Unternehmens Simulators
USIMO2

ARBEITSPLAN

PLANUNGS-ZEITPUNKT	IN STUNDEN	2880	IN TAGEN	120	
	IN WOCHEN	24	IN MONATEN	6	
SIMULATIONS-ZEITPKT	IN STUNDEN	2880	IN TAGEN	120	(VOLLENDET)
	IN WOCHEN	24	IN MONATEN	6	(VOLLENDET)

BAU- TEIL	ARB- GANG	STUE- ZEIT	RUE- ZEIT	BETR.- MITTL	ARBTs KRAFT	BAU TL	ANZ	BAU TL	ANZ	BAU TL	ANZ	BAU TL	ANZ	BAU TL	ANZ		
100	1	3	28	4	5	101	1	0	0	0	0	0	0	0	0	0	1
101	2	4	23	1	6	102	1	0	0	0	0	0	0	0	0	0	2
102	3	2	10	8	5	120	1	110	1	150	2	0	0	0	0	0	3
120	4	2	26	6	4	130	2	140	3	0	0	0	0	0	0	0	4
110	4	5	17	1	1	503	1	0	0	0	0	0	0	0	0	0	5
150	4	4	18	4	2	151	1	0	0	0	0	0	0	0	0	0	6
130	5	2	21	2	4	131	1	0	0	0	0	0	0	0	0	0	7
131	6	5	10	6	5	132	1	0	0	0	0	0	0	0	0	0	8
132	7	4	23	1	3	501	1	0	0	0	0	0	0	0	0	0	9
140	5	6	13	3	2	502	1	0	0	0	0	0	0	0	0	0	10
151	5	4	29	7	5	160	2	170	2	0	0	0	0	0	0	0	11
160	6	6	10	6	3	161	1	0	0	0	0	0	0	0	0	0	12
161	7	6	24	7	2	162	1	0	0	0	0	0	0	0	0	0	13
162	8	4	19	7	4	504	1	0	0	0	0	0	0	0	0	0	14
170	6	3	20	7	5	171	1	0	0	0	0	0	0	0	0	0	15
171	7	5	27	6	3	505	1	0	0	0	0	0	0	0	0	0	16
200	1	4	14	3	6	201	1	0	0	0	0	0	0	0	0	0	17
201	2	6	11	4	3	210	2	260	2	270	2	0	0	0	0	0	18
210	3	4	9	5	1	211	1	0	0	0	0	0	0	0	0	0	19
260	3	4	26	3	6	504	2	0	0	0	0	0	0	0	0	0	20
270	3	6	14	4	3	271	1	0	0	0	0	0	0	0	0	0	21
211	4	5	25	7	3	230	2	240	1	250	3	0	0	0	0	0	22
271	4	6	22	6	1	280	2	290	4	0	0	0	0	0	0	0	23
230	5	6	28	5	5	231	1	0	0	0	0	0	0	0	0	0	24
231	6	6	21	6	2	502	1	0	0	0	0	0	0	0	0	0	25
240	5	4	21	6	1	503	1	0	0	0	0	0	0	0	0	0	26
250	5	3	22	8	6	251	1	0	0	0	0	0	0	0	0	0	27
251	6	4	19	2	5	252	1	0	0	0	0	0	0	0	0	0	28
252	7	6	11	1	4	505	1	0	0	0	0	0	0	0	0	0	29
280	5	6	17	6	2	506	1	0	0	0	0	0	0	0	0	0	30
290	5	4	12	2	2	291	1	0	0	0	0	0	0	0	0	0	31
291	6	3	16	8	4	501	1	0	0	0	0	0	0	0	0	0	32
300	1	4	19	3	1	301	1	0	0	0	0	0	0	0	0	0	33
301	2	2	30	7	2	310	2	320	3	330	2	0	0	0	0	0	34
310	3	5	16	8	2	311	1	0	0	0	0	0	0	0	0	0	35
320	3	4	14	4	2	502	1	0	0	0	0	0	0	0	0	0	36
330	3	4	12	4	5	370	2	380	3	390	4	0	0	0	0	0	37
311	4	4	27	6	5	340	3	350	2	360	4	0	0	0	0	0	38
370	4	5	12	2	5	371	1	0	0	0	0	0	0	0	0	0	39
371	5	4	9	8	5	372	1	0	0	0	0	0	0	0	0	0	40
372	6	4	22	8	3	501	1	0	0	0	0	0	0	0	0	0	41
380	4	6	17	8	1	381	1	0	0	0	0	0	0	0	0	0	42
381	5	6	11	1	2	503	1	0	0	0	0	0	0	0	0	0	43
390	4	6	26	4	6	391	1	0	0	0	0	0	0	0	0	0	44
391	5	4	25	6	3	392	1	0	0	0	0	0	0	0	0	0	45
392	6	6	10	6	5	506	1	0	0	0	0	0	0	0	0	0	46
340	5	2	14	6	3	341	1	0	0	0	0	0	0	0	0	0	47
341	6	4	29	2	5	342	1	0	0	0	0	0	0	0	0	0	48
342	7	6	25	5	6	501	1	0	0	0	0	0	0	0	0	0	49
350	5	5	26	4	4	351	1	0	0	0	0	0	0	0	0	0	50
351	6	4	18	3	5	504	1	0	0	0	0	0	0	0	0	0	51
360	5	5	13	5	1	505	1	0	0	0	0	0	0	0	0	0	52
400	1	6	27	2	2	401	1	0	0	0	0	0	0	0	0	0	53
401	2	5	27	4	3	410	1	420	3	430	2	440	4	0	0	0	54
410	3	6	18	2	4	411	1	0	0	0	0	0	0	0	0	0	55
411	4	3	24	6	5	412	1	0	0	0	0	0	0	0	0	0	56
412	5	4	11	4	6	450	1	460	2	0	0	0	0	0	0	0	57
420	3	2	25	4	2	503	1	0	0	0	0	0	0	0	0	0	58
430	3	4	15	4	3	431	1	0	0	0	0	0	0	0	0	0	59
431	4	4	26	6	4	504	1	0	0	0	0	0	0	0	0	0	60
440	3	5	10	8	4	470	2	480	3	490	2	0	0	0	0	0	61
450	6	5	25	1	1	451	1	0	0	0	0	0	0	0	0	0	62
451	7	5	10	5	3	452	1	0	0	0	0	0	0	0	0	0	63
452	8	6	15	2	6	501	1	0	0	0	0	0	0	0	0	0	64
460	6	4	12	8	4	502	1	0	0	0	0	0	0	0	0	0	65
470	4	4	12	1	3	471	1	0	0	0	0	0	0	0	0	0	66
471	5	5	22	2	1	472	1	0	0	0	0	0	0	0	0	0	67
472	6	2	18	6	1	505	1	0	0	0	0	0	0	0	0	0	68
480	4	6	20	4	6	481	1	0	0	0	0	0	0	0	0	0	69
490	4	5	10	6	6	491	1	0	0	0	0	0	0	0	0	0	70
491	5	4	21	4	4	492	1	0	0	0	0	0	0	0	0	0	71
481	5	4	17	6	4	506	1	0	0	0	0	0	0	0	0	0	72
492	6	6	17	4	4	507	1	0	0	0	0	0	0	0	0	0	73

B E D A R F S P L A N

PLANUNGS-ZEITPUNKT IN STUNDEN 2000 IN TAGEN 120
 24 IN MONATEN 6
 SIMULATIONS-ZEITPUNKT IN STUNDEN 2912 IN TAGEN 121 (VOLLSTÄNDIG)
 24 IN MONATEN 6 (VOLLSTÄNDIG)
 ABRAUS-ZEITPUNKT IN STUNDEN 135
 PLANUNGSVORLAUF IN WOCHE EFFIZIENZ 5
 VORBEREITETER PLANUNGSVORLAUF 3 WOCHE

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528	529	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550	551	552	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571	572	573	574	575	576	577	578	579	580	581	582	583	584	585	586	587	588	589	590	591	592	593	594	595	596	597	598	599	600	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620	621	622	623	624	625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640	641	642	643	644	645	646	647	648	649	650	651	652	653	654	655	656	657	658	659	660	661	662	663	664	665	666	667	668	669	670	671	672	673	674	675	676	677	678	679	680	681	682	683	684	685	686	687	688	689	690	691	692	693	694	695	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711	712	713	714	715	716	717	718	719	720	721	722	723	724	725	726	727	728	729	730	731	732	733	734	735	736	737	738	739	740	741	742	743	744	745	746	747	748	749	750	751	752	753	754	755	756	757	758	759	760	761	762	763	764	765	766	767	768	769	770	771	772	773	774	775	776	777	778	779	780	781	782	783	784	785	786	787	788	789	790	791	792	793	794	795	796	797	798	799	800	801	802	803	804	805	806	807	808	809	810	811	812	813	814	815	816	817	818	819	820	821	822	823	824	825	826	827	828	829	830	831	832	833	834	835	836	837	838	839	840	841	842	843	844	845	846	847	848	849	850	851	852	853	854	855	856	857	858	859	860	861	862	863	864	865	866	867	868	869	870	871	872	873	874	875	876	877	878	879	880	881	882	883	884	885	886	887	888	889	890	891	892	893	894	895	896	897	898	899	900	901	902	903	904	905	906	907	908	909	910	911	912	913	914	915	916	917	918	919	920	921	922	923	924	925	926	927	928	929	930	931	932	933	934	935	936	937	938	939	940	941	942	943	944	945	946	947	948	949	950	951	952	953	954	955	956	957	958	959	960	961	962	963	964	965	966	967	968	969	970	971	972	973	974	975	976	977	978	979	980	981	982	983	984	985	986	987	988	989	990	991	992	993	994	995	996	997	998	999	1000	1001	1002	1003	1004	1005	1006	1007	1008	1009	1010	1011	1012	1013	1014	1015	1016	1017	1018	1019	1020	1021	1022	1023	1024	1025	1026	1027	1028	1029	1030	1031	1032	1033	1034	1035	1036	1037	1038	1039	1040	1041	1042	1043	1044	1045	1046	1047	1048	1049	1050	1051	1052	1053	1054	1055	1056	1057	1058	1059	1060	1061	1062	1063	1064	1065	1066	1067	1068	1069	1070	1071	1072	1073	1074	1075	1076	1077	1078	1079	1080	1081	1082	1083	1084	1085	1086	1087	1088	1089	1090	1091	1092	1093	1094	1095	1096	1097	1098	1099	1100	1101	1102	1103	1104	1105	1106	1107	1108	1109	1110	1111	1112	1113	1114	1115	1116	1117	1118	1119	1120	1121	1122	1123	1124	1125	1126	1127	1128	1129	1130	1131	1132	1133	1134	1135	1136	1137	1138	1139	1140	1141	1142	1143	1144	1145	1146	1147	1148	1149	1150	1151	1152	1153	1154	1155	1156	1157	1158	1159	1160	1161	1162	1163	1164	1165	1166	1167	1168	1169	1170	1171	1172	1173	1174	1175	1176	1177	1178	1179	1180	1181	1182	1183	1184	1185	1186	1187	1188	1189	1190	1191	1192	1193	1194	1195	1196	1197	1198	1199	1200	1201	1202	1203	1204	1205	1206	1207	1208	1209	1210	1211	1212	1213	1214	1215	1216	1217	1218	1219	1220	1221	1222	1223	1224	1225	1226	1227	1228	1229	1230	1231	1232	1233	1234	1235	1236	1237	1238	1239	1240	1241	1242	1243	1244	1245	1246	1247	1248	1249	1250	1251	1252	1253	1254	1255	1256	1257	1258	1259	1260	1261	1262	1263	1264	1265	1266	1267	1268	1269	1270	1271	1272	1273	1274	1275	1276	1277	1278	1279	1280	1281	1282	1283	1284	1285	1286	1287	1288	1289	1290	1291	1292	1293	1294	1295	1296	1297	1298	1299	1300	1301	1302	1303	1304	1305	1306	1307	1308	1309	1310	1311	1312	1313	1314	1315	1316	1317	1318	1319	1320	1321	1322	1323	1324	1325	1326	1327	1328	1329	1330	1331	1332	1333	1334	1335	1336	1337	1338	1339	1340	1341	1342	1343	1344	1345	1346	1347	1348	1349	1350	1351	1352	1353	1354	1355	1356	1357	1358	1359	1360	1361	1362	1363	1364	1365	1366	1367	1368	1369	1370	1371	1372	1373	1374	1375	1376	1377	1378	1379	1380	1381	1382	1383	1384	1385	1386
---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------

Bild 6:Schlüsselzahlen zur Kennzeichnung des Zustandes der Aufträge

- 11 Auftrag in Elemente aufgelöst
- 22 Auftrag von eingehenden Bauelementen her fertig zur Überschreibung an den Betrieb
- 33 Auftrag von eingehenden Bauelementen und von der Vorlaufzeit fertig zum Überschreiben an den Betrieb
- 44 Auftrag von eingehenden Bauelementen und von der Vorlaufzeit her fertig zum Überschreiben an Betrieb; aber das notwendige Material fehlt
- 40 dto.; verspätet im Hinblick auf die Durchlaufterminierung
- 55 Auftrag von eingehenden Bauelementen und von der Vorlaufzeit her fertig zum Überschreiben an den Betrieb; aber Betrieb (Speicher) mit wartenden Aufträgen gefüllt
- 50 dto.; verspätet nach Durchlaufterminierung
- 66 Auftrag dem Betrieb überschrieben; wartet vor Betriebsmittel auf Bearbeitung
- 60 dto.; verspätet im Hinblick auf die Durchlaufterminierung
- 77 Auftrag wartete mit höchster Priorität vor Betriebsmittel auf Bearbeitung; aber keine geeignete Arbeitskraftart vorhanden
- 70 dto.; verspätet im Hinblick auf die Durchlaufterminierung
- 88 Auftrag in Bearbeitung
- 80 dto.; verspätet im Hinblick auf die Durchlaufterminierung
- 99 Bearbeitung des Auftrags vollendet

B E T R I E B S M I T T E L - B E L E G U N G

PLANUNGS-ZEITPUNKT	IN STUNDEN	2880	IN TAGEN	120	
	IN WOCHEN	24	IN MONATEN	6	
SIMULATIONS-ZEITPKT	IN STUNDEN	2912	IN TAGEN	121	(VOLLENDET)
	IN WOCHEN	24	IN MONATEN	6	(VOLLENDET)

ART	ABGANG	NUTZ STD	BEARB ENDE	SP 29 BEDPL	AUFTR NO	LFD NO
1	18	8	0	0	0	1
1	18	8	2936	126	112	2
1	30	8	2936	91	94	3
1	30	8	0	0	0	4
1	42	8	0	0	0	5
1	54	8	0	0	0	6
2	18	8	0	0	0	7
2	30	8	0	0	0	8
2	42	8	0	0	0	9
2	42	8	2934	58	72	10
2	54	8	0	0	0	11
2	54	8	0	0	0	12
3	18	8	0	0	0	13
3	30	8	0	0	0	14
3	30	8	0	0	0	15
3	42	8	0	0	0	16
3	54	8	0	0	0	17
3	54	8	0	0	0	18
3	54	8	0	0	0	19
4	18	8	2929	73	83	20
4	18	8	0	0	0	21
4	30	8	2936	161	134	22
4	30	8	0	0	0	23
4	42	8	3030	156	134	24
4	54	8	0	0	0	25
5	18	8	0	0	0	26
5	30	8	0	0	0	27
5	30	8	2929	74	83	28
5	42	8	0	0	0	29
5	54	8	0	0	0	30
6	18	8	2930	38	43	31
6	30	8	2934	50	72	32
6	30	8	2958	93	94	33
6	42	8	2978	145	123	34
6	42	8	2931	27	61	35
7	18	8	0	0	0	36
7	30	8	0	0	0	37
7	30	8	0	0	0	38
7	42	8	0	0	0	39
7	42	8	0	0	0	40
7	54	8	0	0	0	41
8	18	8	0	0	0	42
8	18	8	0	0	0	43
8	30	8	0	0	0	44
8	30	8	0	0	0	45
8	42	8	0	0	0	46
8	42	8	0	0	0	47
8	54	8	0	0	0	48

A R B E I T S K R A E F T E - B E L E G U N G

PLANUNGS-ZEITPUNKT	IN STUNDEN	2880	IN TAGEN	120	
	IN WOCHEN	24	IN MONATEN	6	
SIMULATIONS-ZEITPKT	IN STUNDEN	2912	IN TAGEN	121	(VOLLENDET)
	IN WOCHEN	24	IN MONATEN	6	(VOLLENDET)

ART	NUTZ STD	BEARB ENDE	AUF TRG	ZEILE BEDPLA	LFD NO
1	8	2934	72	50	1
1	8	0	0	0	2
1	8	2929	83	74	3
1	8	0	0	0	4
1	8	0	0	0	5
1	8	0	0	0	6
1	8	0	0	0	7
1	8	0	0	0	8
1	8	0	0	0	9
2	8	0	0	0	10
2	8	2934	72	58	11
2	8	0	0	0	12
2	8	0	0	0	13
2	8	0	0	0	14
2	8	0	0	0	15
2	8	0	0	0	16
2	8	0	0	0	17
2	8	0	0	0	18
2	8	0	0	0	19
3	8	0	0	0	20
3	8	2936	94	91	21
3	8	0	0	0	22
3	8	0	0	0	23
3	8	0	0	0	24
3	8	0	0	0	25
3	8	0	0	0	26
3	8	0	0	0	27
3	8	0	0	0	28
3	8	0	0	0	29
4	8	2936	112	126	30
4	8	2929	83	73	31
4	8	2936	134	161	32
4	8	0	0	0	33
4	8	0	0	0	34
4	8	0	0	0	35
4	8	0	0	0	36
4	8	0	0	0	37
4	8	0	0	0	38
4	8	0	0	0	39
5	8	2931	61	27	40
5	8	0	0	0	41
5	8	2930	43	38	42
5	8	0	0	0	43
5	8	2978	123	145	44
5	8	0	0	0	45
5	8	0	0	0	46
5	8	0	0	0	47
5	8	0	0	0	48
5	8	0	0	0	49
6	8	0	0	0	50
6	8	0	0	0	51
6	8	0	0	0	52
6	8	3030	134	156	53
6	8	0	0	0	54
6	8	0	0	0	55
6	8	2958	94	93	56
6	8	0	0	0	57
6	8	0	0	0	58

The Effects of the Insurees' Decisions on the Insurers' Profit

by Ch. Haehling von Lanzaer, London/Canada

1. INTRODUCTION

Most existing automobile insurance systems are characterized by features such as:

Rating Classes

Merit Rating Structures

Provision for Deductibles.

Rating classes are used in order to classify individual insurees according to one or more demographic characteristics such as age, sex, marital status, type and age of vehicle, profession, territory of operation, etc. An individual moves from rating class j ($j = 1, \dots, J$) into another as the respective characteristics change. The purpose of rating classes is to group insurees according to statistical observations. It is useful to include an artificial rating class $J+1$. This rating class can be considered as a source of new insurees as well as a sink for those who leave the insurance system.

Merit rating implies that the insurance premium to be paid by an insuree is related to his claim experience. An individual is classified into a claim experience (or risk) category n ($n = 1, \dots, N$) and receives premium discounts (bonus) according to a predetermined discount schedule. Filing a claim to the insurance company results in penalties in the form of either reduction or total loss of already accumulated premium discounts. Such penalties can even be in the form of surcharges (malus). In mutual insurance companies a portion of the contribution to underwriting profit¹ is frequently redist-

¹ The contribution to underwriting profit is defined as the difference between a year's premium income and claim expenditure (including reserves).

ributed to the insurees in the form of rebates. These rebates are generally related to the number of successive years of claim-free driving and follow therefore, the merit rating concept. Merit rating structures are found in third party liability, collision, and all perils insurance.

Provision for deductibles is an integral part of many insurance policies. Whenever an insuree claims a loss in excess of the deductible he has to carry a portion of the loss equal to the agreed deductible. For losses less than the deductible the liability of the insurance company is waived. Generally, a number of deductible values D_k ($k = 1, \dots, K-1$) are available to choose from. It is useful to consider "no insurance" as a special case with a deductible equal to infinity and identify it by D_K . In automobile insurance provision for deductibles is found in collision, all perils, and comprehensive insurance.

Merit rating structures and provision for deductibles require various decisions by the insurees. Having caused an accident, an individual has to decide whether to file a claim of the loss to his insurance company. In doing so he must either accept a higher premium in future years resulting from his claim experience or incur the costs of the loss himself in order to maintain his preferred position. Decisions of this nature are required for all policies with merit rating structures. When deductibles are part of the insurance policy an individual has to decide whether to purchase greater protection (i.e. a small deductible) at a higher premium or less protection (i.e. a large deductible) at a smaller premium. Obviously, these decisions should be made on economic grounds. In view of the sequential nature of the problem optimal decisions can be derived using dynamic programming [3], [4] and [5].

Optimal and non-optimal decision making by the insurees is of significance to individual insurance companies and the industry. For an existing automobile insurance system with given merit rating structures and specified deductible arrangements, premium income and claim expenditures are functions of

the decisions made by the individual insurees. The purpose of this paper is to develop a model for determining the contribution to underwriting profit of the insurance industry¹ as a function of the decisions made by the insurees within a given automobile insurance system. The analysis is restricted to liability and collision insurance since they are by far the most important sections of an automobile insurance policy.

2. THE MODEL

Since in most automobile insurance systems third party liability and collision insurance are independent (i.e. a claim in liability does not affect the risk category in collision and vice versa), the effects of decisions made by the insurees in liability and collision insurance can be analyzed separately.

Liability Insurance

In order to determine the contribution to underwriting profit we must first concentrate on the decisions to be made by the insurees.

Since no provision for deductibles is made in liability insurance the decision problem an insuree faces is whether to file a claim whenever he is at fault.² In deciding whether to file a claim, an insuree uses a decision rule of the following general form:

$$(1a) \quad L_t^{(1)} - L_t^{(1)}(j,n) \begin{cases} \geq 0 & \text{Claim} \\ < 0 & \text{Do not claim} \end{cases}$$

¹ With minor modifications the model can also be applied on a company basis.

² Selecting liability limits is not considered as a significant decision problem since the marginal costs of higher limits are negligible.

with $L_t^{(1)}$ representing the actual liability loss in a given period t and $L_t^{(1)}(j,n)$ being the critical accident size for an insuree in rating class j and risk category n . The superscript (1) is used to identify liability insurance. Assuming an individual wishes to make his decision as to minimize his costs he has to determine the optimal value of $L_t^{(1)}(j,n)$ and thus makes (1a) into an optimal decision rule. In view of the merit rating structures the optimal value of $L_t^{(1)}(j,n)$ can only be determined with respect to future developments and future decisions. The derivation of the optimal $L_t^{(1)}(j,n)$ must therefore be formulated as a sequential decision process. As shown in [3] and [5] the optimal value for risk category j , $\hat{L}_t^{(1)}(j,n)$, is a function of the probability of having an accident in future years.¹ Non-optimal decisions can be characterized by non-negative values for $L_t^{(1)}(j,n)$ other than $\hat{L}_t^{(1)}(j,n)$. For the purpose of this paper we restrict non-optimal decisions to values less than $\hat{L}_t^{(1)}(j,n)$, since insurees are unlikely to carry large losses themselves [3]. Optimal and non-optimal claim behavior is, therefore, reflected in the value of $L_t^{(1)}(j,n)$ to be used in decision rule (1a) and can be expressed as a function of the probability of having accidents.

The contribution to underwriting profit in liability insurance in a given period t , $Z_t^{(1)}$, can be expressed by

$$(2a) \quad Z_t^{(1)} = X_t^{(1)} - Y_t^{(1)}$$

with $X_t^{(1)}$ being the premium income and $Y_t^{(1)}$ the claim expenditure in t . The premium income can be written as

$$(3a) \quad X_t^{(1)} = \sum_{j=1}^J \sum_{n=1}^N C_t^{(1)}(j,n) \cdot M_t^{(1)}(j,n)$$

¹ It is assumed that an individual causes at most one accident per year. This assumption is realistic for the vast majority of individuals.

where $C_t^{(1)}(j,n)$ is the premium to be paid in the given period by an individual belonging to rating class j and risk category n . $M_t^{(1)}(j,n)$ represents the number of individuals in that rating class and risk category. $C_t^{(1)}(j,n)$ is given by

$$(4a) \quad C_t^{(1)}(j,n) = B_t^{(1)}(j) \cdot (1 - q_t(n)/100)$$

with $B_t^{(1)}(j)$ being the base premium (100%) in rating class j and $q_t(n)$ the percentage discount given to insurees in risk category n according to the existing merit rating system. $M_t^{(1)}(j,n)$, the number of insurees in rating class j and risk category n , can be considered as the result of flows between rating classes and risk categories from $t-1$ to t adjusted by a birth-death process allowing for individuals to enter (accretion) and others to leave (attrition) the automobile insurance system. A flow between rating classes is the result of changes in the demographic characteristics (e.g. a change in the type of insured vehicle). A movement from one risk category to another reflects an insuree's claim experience and therefore, depends on his probability of having an accident and subsequent claim behavior. The claim behavior is determined by decision rule (1a) using $L_t^{(1)}(j,n)$ which, in turn, is a function of the probability of having accidents. Since it cannot be assumed that all insurees in rating class j and in risk category n are iso-hazardous, it is necessary to classify the insurees into subgroups with the individuals in each subgroup s ($s = 1, \dots, S$) having the same accident causing propensity. $M_t^{(1)}(j,n)$ can then be expressed by

$$(5a) \quad M_t^{(1)}(j,n) = \sum_{s=1}^S M_t^{(1)}(j,n,s)$$

with $M_t^{(1)}(j,n,s)$ being the number of insurees belonging to rating class j , risk category n and subgroup s . Now it becomes necessary to account for

the fact that individuals will experience changes in their accident causing propensity. The aspect of accretion and attrition can be handled through the artificial rating class $J+1$ which new insurees come from (births) and "retiring" insurees go to (deaths). Let $p_t^{(1)}(ij, mn, rs)$ be the probability of a transition from i ($i = 1, \dots, J+1$) into j , from m ($m = 1, \dots, N$) into n , and from r ($r = 1, \dots, S$) into s . $M_t^{(1)}(j, n, s)$ can then be expressed by (6a)

$$(6a) \quad M_t^{(1)}(j, n, s) = \sum_{i=1}^{J+1} \sum_{m=1}^N \sum_{r=1}^S M_{t-1}^{(1)}(i, m, r) \cdot p_{t-1}^{(1)}(ij, mn, rs).$$

The critical step is to derive the stochastic matrix $[p_t^{(1)}(ij, mn, rs)]$. A transition from rating class i to rating class j takes place as the respective demographic characteristics change. $\pi_t^{(1)}(ij|m, r)$ for $i = 1, \dots, J$ represents the probability that an insuree in rating class i moves to rating class j given he belongs to risk category m and subgroup r . $\pi_t^{(1)}(ij|\cdot, r)$ for $i = J+1$ is the probability of a new insuree entering rating class j who belongs to subgroup r . Risk categories are, of course, not defined for individuals not yet insured. A transition from risk category m to risk category n occurs if an insuree causes an accident and files a claim with the insurance company. $p_t^{(1)}(i, m, r)$ represents the accident causing propensity. The decision whether to file a claim is determined by (1a). The key quantity in (1a) is the critical accident size which is a function of the accident causing propensity. This can be reflected by introducing r into $L_t^{(1)}(i, m, r)$. The probability of a claim in t given an accident is therefore

$$L_t^{(1)}(i, m, r) \int_0^{\infty} f_t(L) dL = [1 - F_t^{(1)}(L_t^{(1)}(i, m, r))]$$

with $f_t^{(1)}(L)$ being the density function and $F_t^{(1)}(L)$ the distribution function

of the size of an accident.¹ Finally, we have to account for changes in the accident causing propensity. These changes are the result of driving skills improving, remaining the same or deteriorating. Let $g_r^{(1)}(s)$ be the probability that a randomly selected insuree in subgroup r this period will be in subgroup s next period. The specific form of $g_r^{(1)}(s)$ will be dependent on r but is taken to be independent of rating class and risk category. The transition probability $p_t^{(1)}(ij, mn, rs)$ can then be expressed by

$$(7a) \quad p_t^{(1)}(ij, mn, rs) = \begin{cases} \pi_t^{(1)}(ij|m, r) p_t^{(1)}(i, m, r) [1 - F_t^{(1)}(L_t^{(1)}(i, m, r))] g_r^{(1)}(s) & n=m^* \\ & j=1 \dots J \\ & s=1 \dots S \\ \pi_t^{(1)}(ij|m, r) \{ p_t^{(1)}(i, m, r) \cdot F_t^{(1)}(L_t^{(1)}(i, m, r)) + \\ \quad (1 - p_t^{(1)}(i, m, r)) \} g_r^{(1)}(s) & n=m^{**} \\ & j=1 \dots J \\ & s=1 \dots S \\ 0 & n \neq m^*, m^{**} \\ \pi_t^{(1)}(ij|m, r) g_r^{(1)}(s) & j=J+1 \\ & s=1 \dots S \end{cases}$$

for $i = 1 \dots J$
 $m = 1 \dots N$
 $r = 1 \dots S$

$$p_t^{(1)}(ij, mn, rs) = \begin{cases} \pi_t^{(1)}(ij|\cdot, r) g_r^{(1)}(s) & j=1 \dots J \\ & s=1 \dots S \\ [1 - \sum_{j=1}^J \pi_t^{(1)}(ij|\cdot, r)] g_r^{(1)}(s) & j=J+1 \\ & s=1 \dots S \end{cases}$$

for $i = J+1$
 $r = 1 \dots S$

$m^*[m^{**}]$ is the resulting risk category for filing a claim [not filing a claim] according to the given merit rating structure.

The claim expenditure during period t , $Y_t^{(1)}$, is a function of the expected number of accidents, the insurees' claim behavior, and the expected claim size

¹ The density function $f_t^{(1)}(L)$ is taken to be independent of risk category m and subgroup r . The model, however, can be modified to consider $f_t^{(1)}$ to be a function of m and r .

(including the administrative costs of handling claims). $Y_t^{(1)}$ can therefore be expressed by

$$(8a) \quad Y_t^{(1)} = \sum_{j=1}^J \sum_{n=1}^N \sum_{s=1}^S P_t^{(1)}(j,n,s) [1 - F_t^{(1)}(L_t^{(1)}(j,n,s))] \cdot M_t^{(1)}(j,n,s) \cdot A_t^{(1)}(j,n,s) [1+a].$$

$A_t^{(1)}(j,n,s)$ is the expected size of a claim filed by an individual in rating class j , risk category n , and subgroup s . The parameter, a , represents the claim handling factor and is a fraction of $A_t^{(1)}(j,n,s)$. The claim handling factor covers such expenses as administrative costs, legal fees, etc. $A_t^{(1)}(j,n,s)$ is given by (9a)

$$(9a) \quad A_t^{(1)}(j,n,s) = \int_{L_t^{(1)}(j,n,s)}^{\infty} L \cdot f_t^{(1)}(L | L \geq L_t^{(1)}(j,n,s)) dL.$$

In view of the following relationship

$$L_t^{(1)}(j,n,s) \int_{L_t^{(1)}(j,n,s)}^{\infty} L f_t^{(1)}(L | L \geq L_t^{(1)}(j,n,s)) dL = \frac{1}{[1 - F_t^{(1)}(L_t^{(1)}(j,n,s))]} L_t^{(1)}(j,n,s) \int_{L_t^{(1)}(j,n,s)}^{\infty} L f_t^{(1)}(L) dL$$

(8a) becomes (10a)

$$(10a) \quad Y_t^{(1)} = \sum_{j=1}^J \sum_{n=1}^N \sum_{s=1}^S P_t^{(1)}(j,n,s) M_t^{(1)}(j,n,s) \cdot L_t^{(1)}(j,n,s) \int_{L_t^{(1)}(j,n,s)}^{\infty} L f_t^{(1)}(L) dL (1+a).$$

Using (3a), (5a) and (10a) the contribution to underwriting profit can be written as

$$(11a) \quad Z_t^{(1)} = \sum_{j=1}^J \sum_{n=1}^N \sum_{s=1}^S \left\{ C_t^{(1)}(j,n) - P_t^{(1)}(j,n,s) \cdot L_t^{(1)}(j,n,s) \int_{L_t^{(1)}(j,n,s)}^{\infty} L f_t^{(1)}(L) dL \cdot (1+a) \right\} \cdot M_t^{(1)}(j,n,s).$$

The effects of optimal or non-optimal claim behavior on the contribution to underwriting profit can be determined by (11a) using either optimal or

non-optimal values for $L_t^{(1)}(j,n,s)$. The evaluation of (11a), however, requires the specification of $M_t^{(1)}(j,n,s)$. With stationary transition probabilities, $M_t^{(1)}(j,n,s)$ is the steady state result of a regular Markov-process.¹ For non-stationary transition probabilities the system (6a) and (7a) represents a linear flow model.

Collision Insurance

In a similar way a model has to be developed for collision insurance which determines the contribution to underwriting profit as a function of the insurees' decisions.² As indicated above there exist two decision areas in collision insurance: The individual has to decide on the level of insurance protection (i.e. a small or a large deductible including a deductible of infinity representing no collision insurance) and whether he should file a claim for a collision loss given he had purchased insurance. The second decision can be based on decision rule (1b)

$$(1b) \quad L_t^{(2)} - L_t^{(2)}(j,n,k) \begin{cases} \geq 0 & \text{Claim} \\ < 0 & \text{Do not claim} \end{cases}$$

for $k = 1 \dots K-1$

with $L_t^{(2)}$ being the actual collision loss during t and $L_t^{(2)}(j,n,k)$ the critical accident size for an insuree in rating class j and risk category n who has selected deductible value D_k . The superscript (2) is used to identify collision insurance. Determining the optimal value of $L_t^{(2)}(j,n,k)$ has to be

¹ Clearly, the restriction $\sum_{j,n,s} M_t^{(1)}(j,n,s) = M_t$ must be added with M_t representing the total population.

² The derivation of this model is similar to the model presented for liability insurance.

made in view of future developments and future decisions. Formulating the problem of simultaneously selecting $L_t^{(2)}(j,n,k)$ and D_k as a sequential decision process and solving it by dynamic programming using cost minimization as an optimization criterion¹ results in the optimum values for $L_t^{(2)}(j,n,k)$ and the optimal deductible value D_k . As shown in [4] and [5] the optimum values, $\hat{L}_t^{(2)}(j,n,k)$ and \hat{D}_k , are functions of the accident causing propensity. Non-optimal decisions are characterized by using values other than $\hat{L}_t^{(2)}(j,n,k)$ and \hat{D}_k and can also be expressed as a function of the accident causing propensity.

As above the contribution to underwriting profit in collision insurance during period t , $Z_t^{(2)}$, can be written as

$$(2b) \quad Z_t^{(2)} = X_t^{(2)} - Y_t^{(2)}$$

$X_t^{(2)}$ is given by (3b)

$$(3b) \quad X_t^{(2)} = \sum_{j=1}^J \sum_{k=1}^K \sum_{n=1}^N C_t^{(2)}(j,n,k) M_t^{(2)}(j,n,k)$$

with $C_t^{(2)}(j,n,k)$ being the premium paid during period t by an individual in rating class j and risk category n who had selected deductible value D_k . $C_t^{(2)}(j,n,k)$ can be written as (4b)

$$(4b) \quad C_t^{(2)}(j,n,k) = B_k^{(2)}(j) [1 - q_t(n)/100]$$

with $B_k^{(2)}(j)$ being the base premium (100%) for collision insurance with deductible value D_k and $q_t(n)$ the percentage discount given to an individual in risk category n according to the merit rating system. Obviously, $B_K^{(2)}(\cdot) = 0$.

¹ For discussion of the optimization criterion see [8], [2] and [6].

The number of individuals in rating class j and risk category n with deductible value D_k , $M_t^{(2)}(j,n,k)$, can be considered as the result of flows between rating classes and risk categories from period $t-1$ to t adjusted by accretion and attrition. As in liability insurance it cannot be assumed that all insurees are iso-hazardous which makes it necessary to classify them into subgroups with constant accident causing propensity. Since the selection of deductibles (optimal or non-optimal) can be expressed as a function of the accident causing propensity it is possible to write $M_t^{(2)}(j,n,k)$ in terms of s . Let $H_k(j,n)$ be the set of subgroups that constitutes the range for selecting deductible D_k . $M_t^{(2)}(j,n,k)$ can then be written as

$$(5b) \quad M_t^{(2)}(j,n,k) = \sum_{s \in H_k(j,n)} M_t^{(2)}(j,n,s)$$

with $M_t^{(2)}(j,n,s)$ being the number of individuals in rating class j , risk category n and subgroup s . The derivation of $M_t^{(2)}(j,n,s)$ is based on the same considerations used in the corresponding section for liability insurance. Therefore, $M_t^{(2)}(j,n,s)$ is given by (6b)

$$(6b) \quad M_t^{(2)}(j,n,s) = \sum_{i=1}^{J+1} \sum_{m=1}^N \sum_{r=1}^S M_{t-1}^{(2)}(i,m,r) p_{t-1}^{(2)}(ij,mn,rs) .$$

The derivation of the transition probability $p_t^{(2)}(ij,mn,rs)$ is analogous to (7a) and given in (7b). It should be noted, however, that each deductible k is associated with some subgroups via the set $H_k(j,n)$. The critical accident size $L_t^{(2)}(j,n,k)$ can therefore be written as $L_t^{(2)}(j,n,s)$ which implies according to $H_k(j,n)$ the corresponding k .

$$p_t^{(2)}(ij, mn, rs) = \begin{cases} \pi_t^{(2)}(ij|m, r) p_t^{(2)}(i, m, r) \cdot [1 - F_t^{(2)}(L_t^{(2)}(i, m, r))] g_r^{(2)}(s) & n=m^* \\ & j=1 \dots J \\ & s=1 \dots S \\ \pi_t^{(2)}(ij|m, r) \{ p_t^{(2)}(i, m, r) \cdot F_t^{(2)}(L_t^{(2)}(i, m, r)) + \\ \quad (1 - p_t^{(2)}(i, m, r)) \} g_r^{(2)}(s) & n=m^{**} \\ & j=1 \dots J \\ & s=1 \dots S \\ 0 & n \neq m^*, m^{**} \\ \pi_t^{(2)}(ij|m, r) g_r^{(2)}(s) & j=J+1 \\ & s=1 \dots S \end{cases}$$

(7b)

$$\begin{aligned} \text{for } i &= 1 \dots J \\ m &= 1 \dots N \\ r &= 1 \dots S \end{aligned}$$

$$p_t^{(2)}(ij, mn, rs) = \begin{cases} \pi_t^{(2)}(ij|\cdot, r) g_r^{(2)}(s) & j=1 \dots J \\ & s=1 \dots S \\ [1 - \sum_{j=1}^J \pi_t^{(2)}(ij|\cdot, r)] g_r^{(2)}(s) & j=J+1 \\ & s=1 \dots S \end{cases}$$

$$\begin{aligned} \text{for } i &= J+1 \\ r &= 1 \dots S \end{aligned}$$

$m^*[m^{**}]$ again is the resulting risk category for filing a claim [not filing a claim] according to the given merit rating structure.

The claim expenditure in collision insurance during period t , $Y_t^{(2)}$, is analogous to liability insurance and can be expressed by

$$(8b) \quad Y_t^{(2)} = \sum_{j=1}^J \sum_{n=1}^N \sum_{s=1}^S p_t^{(2)}(j, n, s) [1 - F_t^{(2)}(L_t^{(2)}(j, n, s))] M_t^{(2)}(j, n, s) A_t^{(2)}(j, n, s) [1+b]$$

with $A_t^{(2)}(j, n, s)$ being the expected size of a claim filed by an insured in rating class j , risk category n and subgroup s . b is a fraction of $A_t^{(2)}(j, n, s)$ and represents claim handling costs. In view of the following relationship

$$(9b) \quad A_t^{(2)}(j, n, s) = \int_{L_t^{(2)}(j, n, s)}^{\infty} L f_t^{(2)}(L | L \geq L_t^{(2)}(j, n, s)) dL$$

(8b) becomes (10b)

$$(10b) \quad Y_t^{(2)} = \sum_{j=1}^J \sum_{n=1}^N \sum_{s=1}^S p_t^{(2)}(j, n, s) M_t^{(2)}(j, n, s) \int_{L_t^{(2)}(j, n, s)}^{\infty} L f_t^{(2)}(L) dL (1+b)$$

Using (3b), (5b) and (10b) the contribution to underwriting profit in collision insurance can be expressed by

$$(11b) \quad Z_t^{(2)} = \sum_{j=1}^J \sum_{n=1}^N \left\{ \sum_{k=1}^K C_t^{(2)}(j,n,k) \left[\sum_{s \in H_k(j,n)} M_t^{(2)}(j,n,s) \right] - \sum_{s=1}^S P_t^{(2)}(j,n,s) \cdot L_t^{(2)}(j,n,s) \int_0^\infty L f_t^{(2)}(L) dL (1+b) M_t^{(2)}(j,n,s) \right\}$$

The effects of optimal or non-optimal decisions made by the insurees on the contribution to underwriting profit can be determined by (11b) using either optimal or non-optimal values for $L_t^{(2)}(j,n,s)$ and for $H_k(j,n)$. As in liability insurance specification of $M_t^{(2)}(j,n,s)$ is necessary for the evaluation of (11b). With stationary transition probabilities, $M_t^{(2)}(j,n,s)$ is the steady state result of a regular Markov process.¹ For non-stationary transition probabilities the system (6b) and (7b) represents a linear flow model.

3. SAMPLE CALCULATIONS

The approach developed has been applied to evaluate optimal and non-optimal decision making in liability insurance within the German Automobile Insurance System as it existed in 1967. Decision making in collision insurance is not to be evaluated for the following reasons. First, the observed population in various risk categories with a given deductible value is very small. Estimates of the density function of the accident size and of the claim frequency are therefore questionable. Second, collision insurance policies are frequently required if the purchase of the automobile is financed. Since no information is available regarding this aspect any assumptions would be purely conjectural.

¹ As above the restriction $\sum_{j,n,s} M_t^{(2)}(j,n,s) = M_t$ must be added with M_t representing the total population.

The analysis is restricted to one rating class representing private passenger cars (91-115 horsepower) operating in all territories.¹ Flows into and out of this rating class are accounted for by the artificial rating class J+1. The merit rating system is composed of four risk categories referring to claim experiences of 0, 1, 2 and 3 or more years of claim free driving. Table 1 specifies the premium discount given in each risk category and the resulting risk categories in the following year for filing and not filing a claim.

TABLE 1
The Merit Rating System

Risk Category n	Discount q(n)/100	Resulting Risk Category	
		Filing a Claim	Not Filing a Claim
1	0	1	2
2	.1	1	3
3	.3	2	4
4	.5	3	4

Since liability insurance is mandatory in the German Automobile Insurance System the only decision problem the insurees face is whether to file a claim for an accident. These decisions are characterized by the value of $L_t^{(1)}(\cdot, n, s)$ used in decision rule (1a). Typical non-optimal decisions [3] may take on the form of (12a)

$$(12a) \quad L_t^{(1)}(\cdot, n, s) = C_{t+1}^{(1)}(T_1(\cdot, n, s)) - C_{t+1}^{(1)}(T_0(\cdot, n, s))$$

with $T_1[T_0]$ representing a transition operator specifying the risk category for the following year after filing a claim [not filing a claim]. The

¹ Civil servants have been eliminated from that rating class since they are granted an a priori premium discount.

critical accident size according to (12a) is the first year difference between the insurance premiums for filing and not filing a claim.¹ Optimal values for $L_t^{(1)}(\cdot, n, s)$ can be obtained by applying the method suggested in [3] and [5] which requires information regarding the base premium (100%) and the density function $f_t^{(1)}(L)$. The base premium in 1967 was DM 608.00. The only information relating to the density function is the average size of a claim in risk category n , $w_t^{(1)}(n)$, excluding claim handling costs [1]. $w_t^{(1)}(n)$ is related to the density function $f_t^{(1)}(L)$ via

$$(13a) \quad w_t^{(1)}(n) = \int_{L_t^{(1)}(\cdot, n, s)}^{\infty} L f_t^{(1)}(L | L \geq L_t^{(1)}(\cdot, n, s)) dL$$

If one assumes that the observed claim size is the result of a claim behavior according to (12a) and that the density function $f_t^{(1)}(L)$ is exponential with parameter $\lambda_t^{(1)}$, it is possible to determine $\lambda_t^{(1)}$ by evaluating (13a) for each risk category. Since the density function $f_t^{(1)}L$ is taken to be independent of the risk categories, $\lambda_t^{(1)}$ is the weighted average of the values obtained for each risk category. The resulting value for $\lambda_t^{(1)}$ is .000748.² The optimal values for the critical accident size and the values for non-optimal decision according to (12a) are illustrated in Figure 1.

¹ Other non-optimal claim behavior may materialize in a critical accident size equal to zero or some positive constant.

² The density function of the size of an accident and the base premium are taken to be time independent (stationary). This assumption is justified since both are highly correlated.

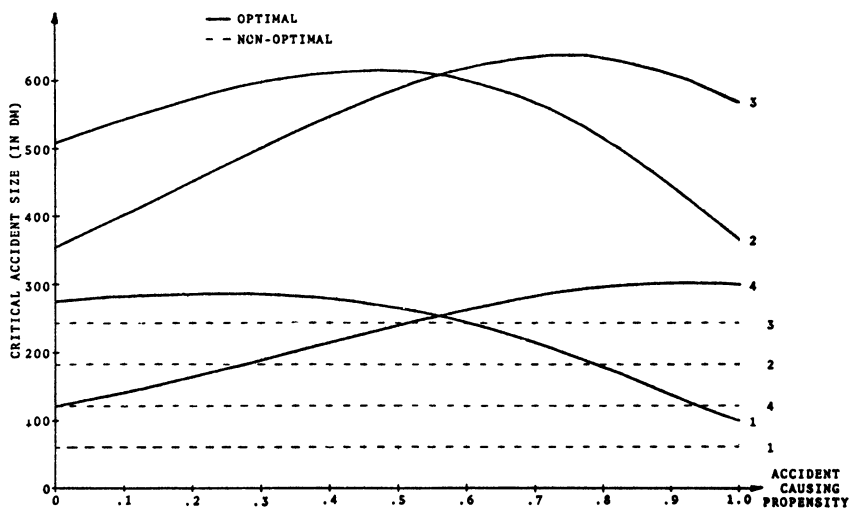


FIGURE 1

The Critical Accident Size for Risk Categories 1 through 4

The evaluation of (11a) requires an estimate of $M_t^{(1)}(\cdot, n, s)$ and this estimation can be made provided the transition probabilities (7a) are given. The transition probabilities relate to

flows between rating classes

transitions between risk categories

changes in the accident causing propensity.

The transitions between risk categories are the result of optimal or non-optimal claim behavior and can easily be formulated. It is more difficult to determine the flows in and out of the rating class considered and the changes in the accident causing propensity. While no information is available with respect to the latter, only the aggregate net result of accretion and attrition is given in [1]. Extensive simulation experiments have been carried out to produce, in connection with the claim behavior (12a), accretion

and attrition probabilities and probabilities relating to changes in the accident causing propensity which are consistent with the reported figures in [1]. No claim is made that these estimates represent the "true" values though they prove to be accurate on a predictive basis (Table 2).

TABLE 2
Actual and Predicted Figures for 1969

Risk Category	Population		Average Claim Frequency	
	Actual	Predicted	Actual	Predicted
1	41,477	42,267	.426	.427
2	24,628	24,033	.292	.296
3	39,333	40,153	.215	.215
44	223,024	222,289	.139	.140
Total	328,462	328,742	.196	.198

The effects of alternative claim behavior on the contribution to underwriting profit can now be evaluated.¹ The claim behavior analysed is optimal decisions (case A), decisions according to (12a) (case B), and decisions to claim every accident (case C). The results are given in Table 3. The figures represent the contribution to underwriting profit in the 10th year of operation under the respective claim behavior with a claim handling factor $a = .30$

¹ The computations have been carried out by Research Assistant William N. Lundberg in the Computing Centre of the School of Business Administration, The University of Western Ontario.

TABLE 3

The Contribution to Underwriting Profit for Alternative
Claim Behavior (DM in Thousands)

Risk Category	Claim Behavior	A	B	C
1		-39,449	-45,660	-46,996
2		+ 170	- 3,026	- 1,857
3		+ 2,659	+ 1,357	+ 3,420
4		+71,120	+77,841	+81,133
Total		+34,500	+30,512	+35,700

According to Table 3 the industry would clearly benefit if the insurees would change from claim behavior according to (12a) (case B) to optimal claim behavior (case A).¹ Similarly, decisions to file a claim for every accident would also improve the industry's position.

The claim handling factor, a , is of critical importance in determining the level of the contribution to underwriting profit. Figure 2 illustrates these relationships. It is interesting to note that for values of $a \geq .382$ optimal claim behavior produces better results than decisions to file a claim for every accident.

¹ The implementation of optimal claim behavior may prove to be difficult since an individual has to estimate his accident causing propensity. A critical accident size, independent of the accident causing propensity but close to the optimal level, seems to be more feasible. Such a near-optimal claim behavior is also desirable from the industry's point of view. (See case E in Figure 3).

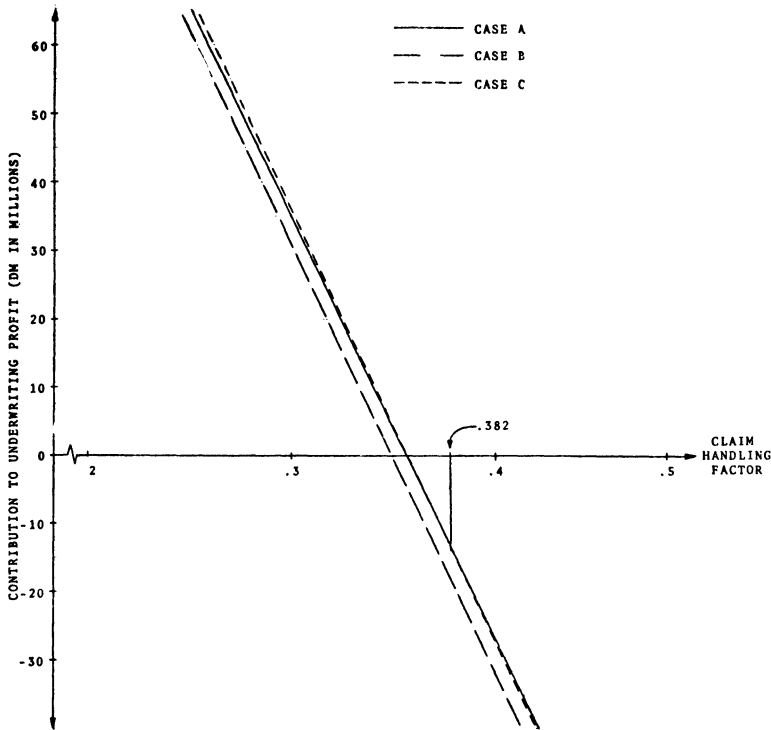


FIGURE 2

The Relationship between the Claim Handling Factor and
the Contribution to Underwriting Profit

Inspection of Figure 2 reveals that non-optimal claim behavior according to (12a) produces consistently poorer results than the two other cases. This was surprising since the critical accident size in case B falls between zero (case C) and the optimal value (case A) and, therefore, a monotonic increasing or decreasing relationship was expected. The results were also derived for claim behavior with a critical accident size of half and twice the size of case B [case D and case E respectively]. Figure 3 illustrates the situation.

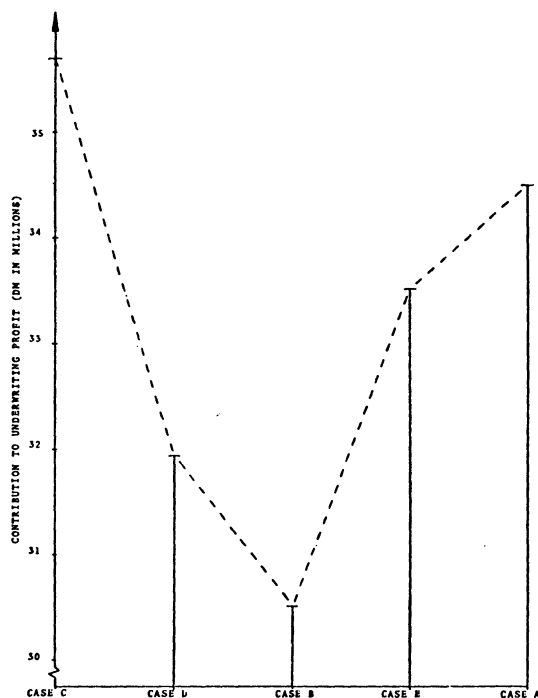


FIGURE 3

The Relationship between Alternative Claim Behavior and the
Contribution to Underwriting Profit

The results reported so far referred to the quasi steady state, a situation which might never be realized. Thus, the contribution to underwriting profit (totals only) is given for the first five years for alternative claim behavior A, B, and C. (Table 4). As can be seen from Table 4 the industry benefits immediately from optimal decisions made by the insureds.

Finally, it is interesting to determine the contribution to underwriting profit per insured in the various risk categories for alternative claim behavior (Table 5).

TABLE 4

The Contribution to Underwriting Profit for Alternative Claim Behavior¹ (DM in Thousands)

Period	A	B	C
1	+ 589	-1,449	-1,973
2	+1,550	- 133	+ 41
3	+2,814	+1,497	+2,165
4	+4,707	+3,534	+4,640
5	+7,290	+6,015	+7,550

TABLE 5

The Contribution to Underwriting Profit per Insuree for Alternative Claim Behavior¹ (in DM)

Risk Category	A	B	C
1	-164.81	-169.49	-164.56
2	+ 1.27	- 21.16	- 11.94
3	+ 10.05	+ 5.21	+ 12.58
4	+ 43.23	+ 48.35	+ 51.68
Average	+ 15.11	+ 13.36	+ 15.64

In all instances the contribution to underwriting profit in risk categories 3 and 4 is positive while it is negative in risk categories 1 and 2. Obviously, risk categories 1 and 2 are subsidized by risk categories 3 and 4. The degree of subsidy, however, varies significantly for alternative claim behavior. The figures in Table 6 are the ratio of the contribution to under-

¹ The claim handling factor, a , is equal to .30.

writing profit per insuree and the average contribution to underwriting profit, e.g. $43.23/15.11 = 2.86$.

TABLE 6

The Degree of Subsidy for Alternative Claim Behavior

Risk Category	A	B	C
3	.67	.39	.80
4	2.86	3.62	3.30

As can be seen from the ratios in Table 6 the degree of subsidy is smallest for optimal claim behavior (case A). Optimal claim behavior, therefore, results also in a more equitable system.¹

4. CONCLUSION

The purpose of this paper was to develop a model by which the insurees' decision making can be evaluated with respect to the contribution to underwriting profit. The model has been illustrated by sample calculations in liability insurance within the German Automobile Insurance System. The analysis indicated that optimal decision making by the insurees is in all instances beneficial for the industry. Naturally, optimal or near optimal decision making can only be realized through the advice of the industry or its agents. This would represent an additional service to the customers. Furthermore it has been shown that optimal claim behavior results in a more equitable system. The models developed can also be used to evaluate the merits of alternative structures of automobile insurance systems before they are implemented and, thus, can contribute to the development of more adequate automobile insurance systems. This aspect is of interest in the ongoing discussion with respect to restructuring automobile insurance, e.g. the introduction of no fault systems.

¹ The above analysis is based on $\lambda^{(1)} = .000748$. A sensitivity analysis using three different values of $\lambda^{(1)}$ produced very similar results.

REFERENCES

- [1] Gesamtstatistik der Kraftfahrtversicherung, 1965-1969.
- [2] Gould, J.P., The Expected Utility Hypothesis and the Selection of Optimal Deductibles for a Given Insurance Policy, Journal of Business, Vol. 42, 2, 1969.
- [3] Haehling von Lanzenauer, C., Entscheidungsregeln für das optimale Verhalten in der Kraftfahrzeug-Versicherung, Unternehmensforschung, Vol. 13, 3, 1969.
- [4] Haehling von Lanzenauer, C., Zur Problematik der Fahrzeugvollversicherung, Unternehmensforschung, Vol. 15, 3, 1971.
- [5] Haehling von Lanzenauer, C., Decision Problems in the Canadian Automobile Insurance System, Journal of Risk and Insurance, Vol. 39, 1, 1972.
- [6] Haehling von Lanzenauer, C., The Expected Cost Hypothesis and the Selection of an Optimal Deductible for a Given Insurance Policy, Journal of Business, Vol. 44, 3, 1971.
- [7] Mehring, J., Die Schadenstruktur in der Kraftfahrt-Haftpflichtversicherung von Personenwagen, Blätter der Deutschen Gesellschaft für Versicherungsmathematik. Vol. 6, 1, 1962.
- [8] Pashigian, B.P., L.L. Schkade, G.M. Menefee, The Selection of an Optimal Deductible for a Given Insurance Policy, Journal of Business, Vol. 39, 1, 1966.
- [9] Seal, H.L., Stochastic Theory of Risk Business, Wiley, New York, 1969.

**A Decision Theoretic Approach to the Design and Analysis
of Industrial Experiments — An Application**

by C.-Y. Lin, Los Angeles

A B S T R A C T

Many of the problems within the domain of classical statistics have been re-examined in recent years from the point of view of Bayesian decision theory. However, a particular area of classical statistics--the design and analysis of experiments--has not been treated in any depth from this viewpoint. The purpose of the paper is to bring decision theory to bear on this class of problems. Specifically, we shall extend previous studies of experimental design models from a Bayesian viewpoint, and focus on cost-effectiveness analysis. In such an analysis, we shall consider explicitly: (a) the probability measure of uncertainty, (b) the economic consequences associated with the decision problem, and (c) the cost and the potential value of the experiment to determine if such an experiment should be conducted to reduce the uncertainty. To demonstrate how this analysis can be used to design and analyze industrial experiments and to aid managers in making decisions, we shall present an actually observed industrial problem.

1. INTRODUCTION

Many of the problems within the domain of classical statistics have been re-examined in recent years from the point of view of Bayesian decision theory. However, a particular area of classical statistics--the design and analysis of experiments--has not been treated in any depth from this viewpoint. The purpose of the paper is to bring decision theory to bear on this class of problems.

Traditionally, experimental design problems have been treated by the analysis of variance techniques, and there is a vast literature devoted to these techniques (see [5]* and the bibliography in it). Recently, however, a few researchers have begun to analyze experimental design problems from a Bayesian viewpoint. For example, Tiao and Tan [7,8] discuss Bayesian inference of a one-way random-effect model, $\tilde{y}_{ij} = \mu + \tilde{a}_i + \tilde{e}_{ij}$. Under the usual normality and independence assumptions, they derive the posterior distributions of the variance components, $\sigma^2 = \text{Var}(\tilde{e}_{ij})$, $\sigma_a^2 = \text{Var}(\tilde{a}_i)$, corresponding to a specific prior distribution, and discuss the features of these distributions. Tiao and Box [6] treat, in a similar fashion, a three-component hierarchical design model, $\tilde{y}_{ijk} = \mu + \tilde{a}_i + \tilde{b}_{ij} + \tilde{e}_{ijk}$, and discuss various features of the posterior distributions of the variance components, $\sigma^2 = \text{Var}(\tilde{e}_{ijk})$, $\sigma_b^2 = \text{Var}(\tilde{b}_{ij})$, and $\sigma_a^2 = \text{Var}(\tilde{a}_i)$.

*Numbers in brackets designate references at end of paper.

The primary concern of these papers are the derivations of posterior distributions of variance components and the use of these distributions to resolve two difficulties that arise if one does traditional analysis of variance--(i) the so-called "negative estimated variance" problem, and (ii) the sensitivity of inferences to departures from underlying assumptions.

John Pratt [2], in his presentation to the Econometric Society in 1967, discussed the orthodox procedure of the analysis of variance and the Bayesian counterpart. Again, he was concerned primarily with the probability distributions of variance components and the use of such distributions for making inferences.

The author, in his recently completed dissertation [1], extended previous Bayesian studies of experimental design models, and focused on the treatment of these models from a decision theoretic viewpoint. This paper, based primarily on that research*, will consider explicitly: (a) the probability measure of uncertainty, (b) the economic consequences associated with the decision problem, and (c) the cost and the potential value of the experiment to determine if such an experiment should be conducted to reduce the uncertainty. To demonstrate how the analysis can be used to design and analyze industrial experiments and to aid managers in making decisions, we shall present an actually observed industrial problem.

* The author wishes to express his deep gratitude to Professors Gordon Kaufman (MIT), Harold Freeman (MIT), and Arthur Schleifer, Jr. (Harvard University) for their generous guidance and encouragement during that research.

2. STATEMENT OF THE PROBLEM

A manufacturer uses various cutting tools to cut a large quantity of workpieces. A tool is used until it becomes dull as judged by the operator. The dull tool is then removed from the machine for resharpening and a sharp tool is placed on the machine. The cost for using a specific type of tool to cut a given number of specified workpieces is a function of the number of changings and resharpenings of tools (hereafter called number of tool-setups). The number of tool-setups required, in turn, depends on the cutting conditions, e.g., velocity and depth of cut. Both of these factors may be set at two alternate levels:

Velocity:	600 rpm or 500 rpm
Depth of Cut:	.020 inches or .003 inches

In other words, there are four combinations of levels of factors--sets of cutting conditions. The tool manager, therefore, wishes to: (a) choose the "best" set of cutting conditions from among the four combinations--best in the sense that the expected cost of cutting workpieces is minimized; (b) determine whether or not experimentation is economically worthwhile prior to a decision to choose a particular set of cutting conditions; and (c) if experimentation is warranted, design an experiment which explicitly considers the cost of experimentation and the expected value of the information to be obtained from the experiment.

3. CHOICE OF THE "BEST" COMBINATION OF LEVELS OF FACTORS

Model--Data Generating Process

To begin our analysis of this problem, we let μ_{ij} denote the mean number of tool-setups required for cutting ten work-pieces when cutting conditions are controlled as follows: velocity at i^{th} level, and depth of cut at j^{th} level. Then we have a two-factor model:

$$\tilde{y}_{ijn} = \mu_{ij} + \tilde{\epsilon}_{ijn}, \quad \begin{array}{l} i = 1, 2 \\ j = 1, 2 \\ n = 1, \dots, N_{ij} \end{array} \quad (1)$$

There are four combinations of levels of factors. To simplify our subsequent analysis, we re-label the subscript (ij) as m, where $m=1, \dots, 4$, i.e., (11) as 1, (12) as 2, (21) as 3, and (22) as 4. Then the above two-factor model, each factor having two levels, may be written as the following one-factor model with four levels,

$$\tilde{y}_{mn} = \mu_m + \tilde{\epsilon}_{mn}, \quad \begin{array}{l} m = 1, \dots, 4 \\ n = 1, \dots, N_m \end{array} \quad (2)$$

We may write (2) in matrix notation*,

$$\underline{\tilde{y}} = \underline{\underline{X}} \underline{\mu} + \underline{\tilde{\epsilon}}, \quad \text{where } \underline{\underline{X}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 1 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 1 & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 1 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad (3)$$

* A single underline is used to designate a vector, and a double underline is used to designate a matrix.

In our subsequent analysis, we shall make the usual homoscedasticity and independence assumptions which allow us to express $\text{Var}(\tilde{\epsilon}) = h^{-1} \underline{I}$, where \underline{I} is an identity matrix. We shall further assume h is known.

Economic Structure

Following a discussion with the appropriate people in the plant, we obtained

c_o = cost of operating time = \$.15 per minute,

c_t = tool cost = \$4/tool,

c_s = cost of changing and resharpening a tool = \$5.00 per setup,

t_m = time to machine a workpiece when the m^{th} set of cutting conditions is chosen, $t_m = 1/(\text{velocity} \times \text{feed})$ minutes,

t_h = time to handle (load and unload) a workpiece = 1/3 minute,

l = "tool life" (average number of resharpenings allowed before a tool wears out) = 8 times,

μ_m = mean number of tool setups required for cutting 10 workpieces when the m^{th} set of cutting conditions is used. This is not known with certainty,

d = demand (number of workpieces to be cut in a period of time) = 1000, where 1000 is the anticipated three month demand.

If we let C_m denote the total cost of cutting d workpieces when the m^{th} set of cutting conditions is chosen, then

$$C_m = dc_o(t_h + t_m) + \frac{d}{10} (c_s + c_t/l) \mu_m. \quad (4)$$

Notice that the cutting conditions affect the cost of cutting only through t_m and μ_m . To treat costs as negative values, we let

$$K_m = -dc_o(t_h + t_m), \quad (5a)$$

$$k = -\frac{d}{10} (c_s + c_t/l). \quad (5b)$$

Then, in terms of a value function, (4) becomes

$$v_m = K_m + k\mu_m, \quad m = 1, \dots, 4, \quad (5c)$$

or in matrix notation*,

$$\underline{v} = \underline{K} + k\underline{\mu}. \quad (5d)$$

The values of \underline{K} and k obtained via (5a and 5b) are:

$$\underline{K} = \begin{bmatrix} - \$150 \\ - \$150 \\ - \$170 \\ - \$170 \end{bmatrix}, \quad k = -\$550. \quad (6)$$

If $\underline{\mu}$ were known with certainty we could easily compute \underline{v} in (5) and choose the best set of cutting conditions.

Prior Distribution of $\underline{\mu}$

We elicit an expert's probability judgments about $\underline{\mu}$. The subjective distribution of $\underline{\mu}$ is approximately multivariate normal with mean and variance shown below:

$$\underline{\mu}' = \begin{bmatrix} .80772 \\ .51128 \\ .60786 \\ .42447 \end{bmatrix}, \quad \underline{\underline{\mu}}' = \begin{bmatrix} .05367 & .03578 & .04294 & .02863 \\ .03578 & .03337 & .02863 & .01908 \\ .04294 & .02863 & .05311 & .02290 \\ .02863 & .01908 & .02290 & .02855 \end{bmatrix} \quad (7)$$

* In general, k may be a matrix with k_m , $m = 1, \dots, 4$ on the main diagonal, and zero's elsewhere.

Terminal Analysis

Our objective here is solely to find the set of cutting conditions with the greatest expected value. Using the \underline{K} and k of (6) and the $\underline{\bar{u}}'$ of (7) we compute $\underline{\bar{v}}'$ via equation (5d) and obtain:

$$\underline{\bar{v}}' = \begin{bmatrix} -\$594 \\ -\$431 \\ -\$504 \\ -\$403 \end{bmatrix} \quad (8)$$

Thus, if the tool manager is to make a terminal choice, he clearly should choose the fourth set of cutting conditions (velocity 500 rpm, depth of cut .003 inches).

Expected Value of Perfect Information

As pointed out in [1], the above formulation of the problem is exactly the same as the selection of best of several processes in [3] and [4]. We adopt the convention of numbering the process with the greatest prior expected value as M^{th} (last) process*.

We then define

$$\delta_m = v_m - v_4, \quad m = 1, \dots, 4 \quad (9a)$$

or in matrix notation,

$$\underline{\delta} = \underline{\underline{B}} \underline{v}, \quad \text{where } \underline{\underline{B}} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}. \quad (9b)$$

Then, using the results in the above mentioned references, the expression for evaluating the expected value of perfect informa-

* In our present example, the optimal process is coincidentally the fourth (last) process and hence no re-numbering is needed.

tion for our tool-setup problem is

$$EVPI = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \max \{ \delta_1, \delta_2, \delta_3, 0 \} f_N^{(3)}(\underline{\delta} | \underline{\bar{\delta}}, \underline{\check{\delta}}) d\delta_1 d\delta_2 d\delta_3 \quad (10a)$$

where $\underline{\bar{\delta}} = \underline{B} \underline{K} + \underline{B} \underline{k} \underline{\bar{\mu}}$,

$$\underline{\check{\delta}} = (\underline{B} \underline{k}) \underline{\check{\mu}} (\underline{B} \underline{k})^t = h^{-1} (\underline{B} \underline{k}) \underline{n}^{-1} (\underline{B} \underline{k})^t.$$

Using the \underline{K} and \underline{k} as in (6), and $\underline{\bar{\mu}}$ and $\underline{\check{\mu}}$ as in (7), we obtain

$$\underline{\bar{\delta}} = \begin{bmatrix} -191.0 \\ -27.7 \\ -101.0 \end{bmatrix}, \quad \underline{\check{\delta}} = \begin{bmatrix} 7550 & 5030 & 6040 \\ 5030 & 7190 & 4600 \\ 6040 & 4600 & 10800 \end{bmatrix}. \quad (10b)$$

The quadrature method in [1] was used to evaluate this EVPI and the result is \$25.92.

Cost of Experimentation

An experiment will be conducted as follows: A set of cutting conditions is chosen and a tool is selected and used to cut the specified workpieces. When the tool becomes dull, the number of cuts made, r_{m_1} , is recorded. If $r_{m_1} \geq 10$, then the number of tool-setups required to make ten cuts, y_m , is taken as $y_m = \frac{10}{r_{m_1}}$. This implies that tool life is linear with respect to the number of cuts. If $r_{m_1} < 10$, then another tool is used and the number of cuts made, r_{m_2} , is recorded. If $r_{m_1} + r_{m_2} \geq 10$, then y_m is taken as $y_m = 1 + \frac{10 - r_{m_1}}{r_{m_2}}$. Otherwise, the procedure is repeated until at least ten cuts are observed.

The prior distributions of $\tilde{\mu}_m$, $m = 1, \dots, 4$ are such that $P(\tilde{\mu}_m > 2) = 0$. Therefore, letting $c_m = c_o (t_h + t_m)$ and $c_f = c_s + c_t/l$, we may express the prior expected cost of sampling, k_{sm} , when one observation is to be taken on the m^{th} process, as

$$k_{sm} = c_m \cdot E_O^2\left(\frac{10}{\tilde{\mu}_m}\right) + c_f \cdot P(\tilde{\mu}_m \leq 1) \\ + c_m \cdot \{P(\tilde{\mu}_m > 1) \cdot E_O^2\left(\frac{10}{\tilde{\mu}_m}\right)\} + 2c_f \cdot P(\tilde{\mu}_m > 1). \quad (11a)$$

Using the cost figures and the prior distribution of $\tilde{\mu}_m$, $m=1, \dots, 4$ stated earlier, we obtain:

$$\underline{k}_s = \begin{bmatrix} \$8.89 \\ \$9.05 \\ \$9.30 \\ \$10.32 \end{bmatrix} \quad (11b)$$

If n_m observations are to be taken on the m^{th} process, the total expected cost of sampling, ECS, clearly is

$$ECS = \sum_{m=1}^M n_m k_{sm}. \quad (12)$$

Since the EVPI is only \$25.92 and since a single additional observation on just one of the four processes will cost from \$8.89 to \$10.32 depending on the process, we conjecture that the expected net gain of further experimentation will almost certainly be negative for all possible experiments; at best it can be a negligible positive amount.

EVSI, ENGS, and the Optimal Design of an Experiment

Recall that the data generating process is $\tilde{\underline{y}} = \underline{X}\underline{\mu} + \tilde{\underline{\epsilon}}$, where $\text{Var}(\tilde{\underline{\epsilon}}) = h^{-1}\underline{I}$ is assumed known. To evaluate EVSI for our tool-setup problem, we must know h . The expression for the EVSI is:

$$\text{EVSI} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \max \{ \bar{\delta}_1'', \bar{\delta}_2'', \bar{\delta}_3'', 0 \} f_N^{(3)}(\bar{\delta}'' | \bar{\delta}', \check{\underline{\delta}}'') d\bar{\delta}_1'' d\bar{\delta}_2'' d\bar{\delta}_3'' \quad (1)$$

$$\text{where } \check{\underline{\delta}}'' = (\underline{B} \ \underline{k}) \ \check{\underline{\mu}}'' \ (\underline{B} \ \underline{k})^t = h^{-1}(\underline{B} \ \underline{k}) \ \underline{n}^{*-1}(\underline{B} \ \underline{k})^t,$$

$$\underline{n}^{*-1} = \underline{n}'^{-1} - \underline{n}''^{-1}.$$

From Chapter 3 in [1], Chapter 24 in [3], or Chapter 13 in [4], we have

$$\bar{\underline{\mu}}'' = \underline{n}''^{-1} (\underline{n}' \bar{\underline{\mu}}' + \underline{n} \ \underline{b}). \quad (14)$$

We used the method of hypothetical future samples [9] to elicit the expert's judgments about $\bar{\underline{\mu}}''$. We stated: "We have elicited your probability judgments about the mean numbers of tool-setups required when the tool is used under four different cutting conditions. We would like to know how sampling outcomes affect your judgments. Suppose that we had taken a sample of one on each of the four sets of cutting conditions and observed {1.2, .75, .9, .6}. Now what are your judgments about the mean number of tool-setups required?" We substituted his answer for $\bar{\underline{\mu}}''$ in (14) to compute \underline{n}' , and then used this \underline{n}' , and $\check{\underline{\mu}}'$ in (7) together with the relation $\check{\underline{\mu}}' = (h\underline{n}')^{-1}$ to obtain $h = 10$. Rather than assigning a gamma

distribution to h [Chapter 5B of 4], we performed a sensitivity analysis for h --used three values of h ($h=5,10,15$) to find optimal sample size.

For given h and \underline{n} we may compute $\frac{\underline{v}}{\underline{\delta}}$ and evaluate the EVSI via the quadrature method presented in [1]. The ENGS is simply EVSI less ECS. The optimal sample size n^0 is that which maximizes the ENGS.

The optimal sample sizes obtained via a computer search are zeros for all three values of h , as we conjectured in the last section, based on the relative size of the EVPI and the cost of sampling.

For purposes of illustration, the sampling costs for all four processes are reduced to one-hundredth of the actual figures in order to obtain non-zero optimal sizes. The results are $n_1^0 = .001$, $n_2^0 = 21.1$, $n_3^0 = 12.5$, $n_4^0 = 19.6$.

4. ANALYSIS OF EFFECTS OF LEVELS OF FACTORS ON A RESPONSE VARIABLE

A classical analysis of variance technique applied to our tool-setup problem is to adopt one of the two models below:

$$\tilde{y}_{ijn} = \mu + \rho_i + \kappa_j + \tilde{\epsilon}_{ijn} \text{ (without interactions),} \quad (15)$$

$$\tilde{y}_{ijn} = \mu + \rho_i + \kappa_j + \tau_{ijn} + \tilde{\epsilon}_{ijn} \text{ (with interactions),} \quad (16)$$

and to use the experimental data generated according to the model chosen to test the null hypothesis that the cutting conditions have no effects on the number of tool-setups required to make a given number of cuts. This approach is not particularly useful in our problem, for the parameters to be tested are not of interest to the tool manager. To see this, let us focus on the first model.

As discussed in [1], we may write

$$\mu_{ij} = \mu + \rho_i + \kappa_j, \quad i = 1, 2, \quad j = 1, 2. \quad (17)$$

If we let

$$\mu_{i.} = \frac{1}{2}(\mu_{i1} + \mu_{i2}), \quad i = 1, 2, \quad (18a)$$

$$\mu_{.j} = \frac{1}{2}(\mu_{1j} + \mu_{2j}), \quad j = 1, 2, \quad (18b)$$

$$\mu = \frac{1}{4}(\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22}), \quad (18c)$$

then we have

$$\rho_i = \mu_{i.} - \mu \quad i = 1, 2, \quad (19a)$$

$$\kappa_j = \mu_{.j} - \mu \quad j = 1, 2. \quad (19b)$$

The tool manager is interested in knowing μ_{ij} in order to choose the "best" combination of (ij), and to use the tools at this set of cutting conditions. The averages, $\mu_{i.}$, $\mu_{.j}$, μ , and thus the differential effects, ρ_i , κ_j , are meaningless with respect to the actual cutting operations*. However, if one is

* Here we examine the relevancy of these parameters for this tool-setup problem. See [2] for some general comments on testing hypotheses in linear models.

interested in these parameters, the distribution of these parameters can be derived [1], from the distribution given in equation (7).

5. CONCLUSION

This study has provided an operational and computationally feasible method for treating a class of experimental design problems from a decision theoretic viewpoint, considering explicitly the economic consequences, the cost and the value of experimentation, and the probability distributions of unknown parameters. An application of the methodology to a real-world problem of industrial experimentation has demonstrated the usefulness of this methodology and the inadequacy of the classical analysis of variance techniques.

REFERENCES

1. Lin, Chi-Yuan, "A Decision Theoretic Approach to the Design and Analysis of Industrial Experiments," Doctoral Thesis, Massachusetts Institute of Technology, 1968.
2. Pratt, John, "Testing Hypotheses in Linear Models: Some Comments and Bayesian Alternatives," presented to the Econometric Society, December, 1967.
3. Pratt, John, Raiffa, Howard, and Schlaifer, Robert, Introduction to Statistical Decision Theory, Preliminary Edition, McGraw-Hill, New York, 1965.
4. Raiffa, Howard and Schlaifer, Robert, Applied Statistical Decision Theory, Harvard University, Boston, 1961.
5. Scheffe, Henry, The Analysis of Variance, John Wiley & Sons, New York, 1967.
6. Tiao, G.C., and Box, G.E.P., "Bayesian Analysis of a Three-Component Hierarchical Design Model," Biometrika, Vol. 54, 1967.
7. Tiao, G.C., and Tan, W.Y., "Bayesian Analysis of Random-Effect Models in the Analysis of Variance. I. Posterior Distribution of Variance-Components," Biometrika, Vol. 52, 1965.
8. Tiao, G.C. and Tan, W.Y., "Bayesian Analysis of Random-Effect Models in the Analysis of Variance. II. Effect of Autocorrelated Errors," Biometrika, Vol. 53, 1966.
9. Winkler, R., "The Assessment of Prior Distributions in Bayesian Analysis," Journal of the American Statistical Association, Vol. 62, No. 319, September, 1967.

Zur optimalen Gleitzeitregelung bei stoßweisem Arbeitsanfall
von G. Mensch, Berlin

Weit über 100 Verwaltungen und Betriebe sind in den letzten Jahren zur gleitenden Arbeitszeit übergegangen. Die optimale Regelung der Gleitzeitspannen tritt dabei als betriebswirtschaftliches Entscheidungsproblem auf.

Eine Systemanlage der stoßweise anfallenden Arbeitsmenge, sowie der Betriebsbereitschaft, die sich bei verschiedenen Gleitzeitregelungen verschieden aufbaut, führt zur Ermittlung der optimalen Gleitzeitspanne.

I. Die Gleitzeitregelung im Betrieb

Viele Verwaltungen und Betriebe haben in den letzten Jahren eine neue Arbeitszeitregelung eingeführt und sind zur gleitenden Arbeitszeit übergegangen. Bei einer Arbeitszeitregelung mit starrer Arbeitszeit sind die Mitarbeiter verpflichtet, die Arbeit zu einem bestimmten Zeitpunkt zu beginnen und zu einem bestimmten Zeitpunkt zu beenden. Bei gleitender Arbeitszeit fallen Arbeitsbeginn und Arbeitsschluß in eine Zeitspanne, die Kommt- bzw. die Gehrgleitzeit; die Mitarbeiter bestimmen selbst, wann sie die Tagesarbeit innerhalb dieser Zeitspannen anfangen bzw. beenden,

In zahlreichen Artikeln der Tagespresse und in Fachzeitschriften wurde innerhalb der letzten zwei Jahre von den Erfahrungen mit der gleitenden Arbeitszeit berichtet. Eine systematische Darstellung der arbeitsrechtlichen Vorschriften zu und eine Übersicht über die Praxis dieser jungen Form der Arbeitszeitregelung geben Hackh (1) und Knevels und Zehle (2).

A. Gleitzeitregelung als Entscheidungsproblem

Über die generellen Vorteile einer solchen Regelung mit gleitenden Arbeitszeiten herrscht weitgehende Einmütigkeit zwischen den Tarifpartnern und, in den einzelnen Betrieben, zwischen Management und Betriebs-

rat bzw. Administration und Personalrat. Wo diese Arbeitszeitregelung überhaupt durchführbar ist, dort haben viele Betriebe und ihr Personal davon Nutzen ziehen können. Zahlreiche Aussagen von Managern bestätigen eine Verbesserung der Produktivität. Im allgemeinen wird auch das Betriebsklima günstig beeinflusst, teils als Ausfluß der Kooperation von Betriebsrat und Management bei der Planung der neuen Arbeitszeitregelung, teils wegen der Disponibilitätsvorteile, die die Mitarbeiter genießen. Unter diesen Vorteilen findet sich die Entspannung der Verkehrslage während der rush-hours, aber auch die Möglichkeit, während der Gleitspannen Besorgungen zu machen. Ein wichtiger Vorteil für den Arbeitnehmer besteht darin, daß er Zeitguthaben ansammeln und durch freie Tage ausgleichen kann; dadurch gewinnt er einen größeren Einfluß auf die Bedingungen des Arbeitseinsatzes.

Eine Neuregelung der Arbeitszeit bedarf jedoch gründlicher Planung; ein falsches Modell der gleitenden Arbeitszeit könnte die betrieblichen Vorteile leicht in ihr Gegenteil verkehren, und es könnte die Arbeitnehmer schlechter als bei der alten Regelung stellen. Hier liegt ein betriebswirtschaftliches Entscheidungsproblem vor.

Die gleitende Arbeitszeit kommt für solche Betriebsbereiche nicht infrage, in denen die Anwesenheit der Mitarbeiter während der gesamten - starren - Betriebszeit notwendig ist. Dies ist in der fließenden Fertigung und bei stark verflochtenen Arbeiten meistens der Fall. In solchen Bereichen aber, in denen eine weitgehende Dispositionsfreiheit des Arbeitsablaufs ein Verlagern der Arbeiten auf frühere oder spätere Stunden gestattet, hat die gleitende Arbeitszeit die weiteste Verbreitung gefunden. Während also ein starrer Arbeitsablauf eine Gleitzeitregelung ausschließt, kann ein disponibler Arbeitsanfall mehrere Möglichkeiten einer Gleitzeitregelung zulassen, ohne daß der einen oder anderen Regelung ein deutlicher Vorzug anhaftet. Hier stellt sich die Frage nach der besten Gleitzeitregelung offenbar nicht.

Sie stellt sich dort, wo der Arbeitsanfall flexibel genug ist, um gleitende Arbeitszeiten zu gestatten, aber doch solchen zeitlichen und quantitativen Schwankungen unterliegt, daß eine mangelnde Betriebsbereitschaft - infolge der zeitweiligen Unterbesetzung von Arbeitsplätzen - zu Störungen führt. Hier gilt es eine solche Gleitregelung zu finden, die die Produktivitätsvorteile des Betriebes und die Disponibilitätsvorteile der Mitarbeiter gewähr-

leistet, ohne daß Störungen im Leistungsgefüge eintreten.

Wir gehen von der Voraussetzung aus, daß während der Kernzeit, wenn alle Mitarbeiter anwesend sind, der Arbeitsanfall in etwa mit der Arbeitskapazität übereinstimmt. Dann kommen als Störungsquellen, die einer Einführung der gleitenden Arbeitszeit möglicherweise entgegenstehen könnten, nur eine unvollständige Betriebsbereitschaft und ein starker Arbeitsanfall während der Kommt- und Gehtzeit in Betracht. Die folgende Untersuchung ist dem Fall gewidmet, daß der Arbeitsanfall während der Kommtgleitzeit in den frühen Morgenstunden sehr groß ist. Dann ist es fraglich, ob bei Einführung einer Kommtgleitzeit die Arbeit pünktlich erledigt werden kann, weil die Betriebsbereitschaft sich nur allmählich aufbaut. Ein Musterprofil für den Aufbau und Abbau der Betriebsbereitschaft, gemessen als Prozentanteil anwesender Mitarbeiter, zeigt Abbildung 1.

Ein stoßweiser Arbeitsanfall am Anfang des Arbeitstages kann aus sehr vielen Gründen eintreten. Selbst für schikanöse Gründe finden sich Beispiele (die Praxis des Arztes, der keine Termine vergibt). Liegt ein temporäres

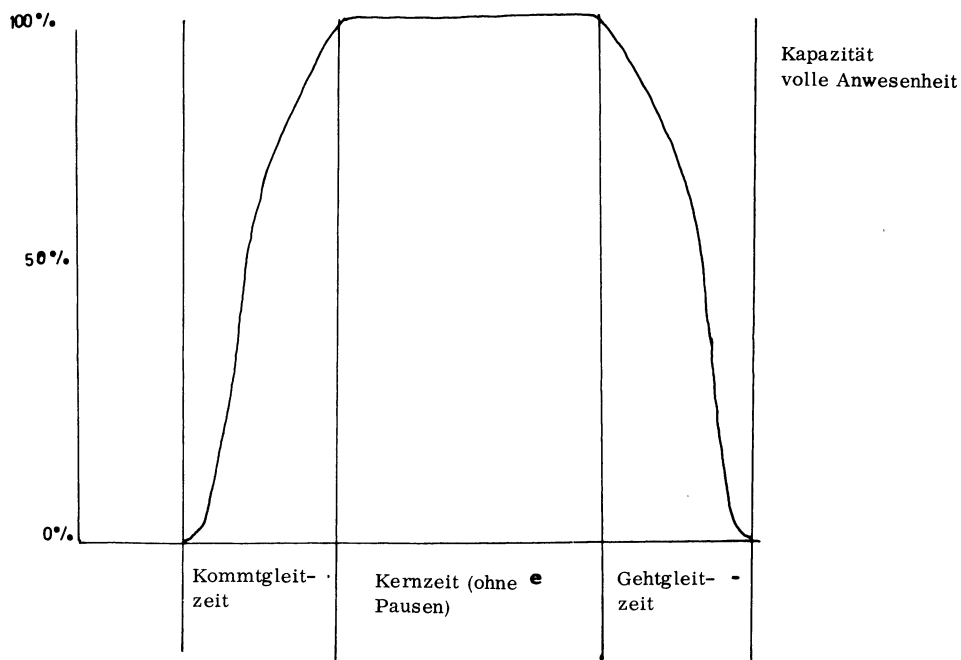


Abbildung 1

Mißverhältnis vor zwischen Arbeitsanfall und Kapazität, dann bleibt häufig einige Arbeit zeitweilig liegen, und zwar meistens in jenen Situationen, wo daraus keine Kosten entstehen (oder, wie beim wartenden Patienten, auf diesen abgewälzt werden). Häufig verursacht die Verzögerung direkte oder indirekte Kosten. Eine solche Situation liegt in vielen Wirtschaftsbereichen vor, etwa beim Handel mit, und bei der Fertigung von schnell verderblichen Gütern. Dazu gehört das Geschäft des Verteilers der Morgenzeitung ebenso wie der Großhandel mit Milchprodukten, die bei Ladenöffnung vor der Tür des Einzelhandels stehen müssen.

Der allgemeinste Fall ist, daß die Kosten oder Opportunitätskosten erst ab Überschreiten einer Schwelle der Toleranz angelastet werden. Das trifft zu für die Verteilung von Telegrammen und Eilbriefen, die über Nacht eintrafen und zwischen 6 und 8 Uhr ausgetragen sein müssen. Diese Problematik wird hier als Fallstudie abgehandelt: Welche Gleitzeitregelung wäre für die Telegramm- und Eilbriefabteilung eines städtischen Postamtes optimal?

B. Ein Gleitzeitmodell

Zur Illustration einer Gleitzeitregelung im öffentlichen Dienst dient das Modell des Landratsamtes Forchheim,

das im November 1969 eingeführt wurde und sich offenbar bewährt hat. Die Mitarbeiter äußerten bei einer Befragung im Mai 1970 fast uneingeschränkte Zustimmung (93,5 % der Befragten), nur 7,5 % der Befragten äußerten ernste Zweifel, aber niemand war absolut gegen die neue Regelung der Arbeitszeit (1, S. 137).

Die Regelung besagt:

6:45 - 8:00	Kommtgleitzeit (Zeitspanne des Erscheinens am Arbeitsplatz)
8:00 - 12:00	1. Kernzeit (alle anwesend)
12:00 - 14:00	gleitende Mittagspause (mindestens 30 Minuten)
14:00 - 16:00	2. Kernzeit (alle anwesend)
16:00 - 18:00	Gehtgleitzeit
(Freitag bis 17:00)	(Zeitspanne des Verlassens des Arbeitsplatzes)

Das Kommen und Gehen der Mitarbeiter während der Gleitzeiten spielte sich bald nach der Einführung der neuen Regelung auf ein deutliches Verhaltensmuster ein, das - nach Hackh (1, S. 142) - auch für andere Belegschaften typisch ist.

Mit einer Streuung von \pm 7 Minuten trafen die Mitarbeiter des Landratsamtes Forchheim zu folgenden Zeitabschnitten innerhalb der Morgengleitzeit, 6:45 bis 8:00, ein:

Saison	6:45 - 7:00 h	7:00 - 7:15 h	7:15 - 7:30 h	7:30 - 8:00 h	tägl. wech- selnd	ins- gesamt
Nov.69- März 70	12,2	15,3	22,6	41,9	8,0	100 %
April70 -Juli70	23,5	15,8	18,1	30,0	12,6	100 %

Quelle: Hackh, a.a.O., S. 142

Tabelle 1

So wie sich viele Mitarbeiter im Frühling/Sommer zum frühen Arbeitsbeginn entschlossen, wählten sie auch ein frühes Arbeitsende. Dagegen kam und ging im Herbst/Winter die große Mehrheit tendenziell später. Dabei wurde offenbar das Kommen und Gehen spontan an den Arbeitsanfall angepaßt. Diese Rücksichtnahme der Mitarbeiter äußerte sich ferner als Großzügigkeit im Gewähren von Zeitkrediten. Rund 40 % der Mitarbeiter sammelten Zeitguthaben von nahezu 5 Stunden pro Monat, weitere 40 % arbeiteten bis zu 10 Stunden im Monat vor. Das saisonbedingte Kommen der Mitarbeiter ist in Abbildung 2 dargestellt.

Das saisonal verschiedene Wachstum der Anwesenheit der Mitarbeiter während der Kommtgleitzeit zeigt Abbildung 3. Die Betriebsbereitschaft, die wir mit der prozentualen

Anwesenheit gleichsetzen, entwickelt sich während der Kommtgleitzeit entsprechend den akkumulierten Häufigkeiten für das Kommen.

Kommt - Verhalten des Personals
in Prozent der Mitarbeiter

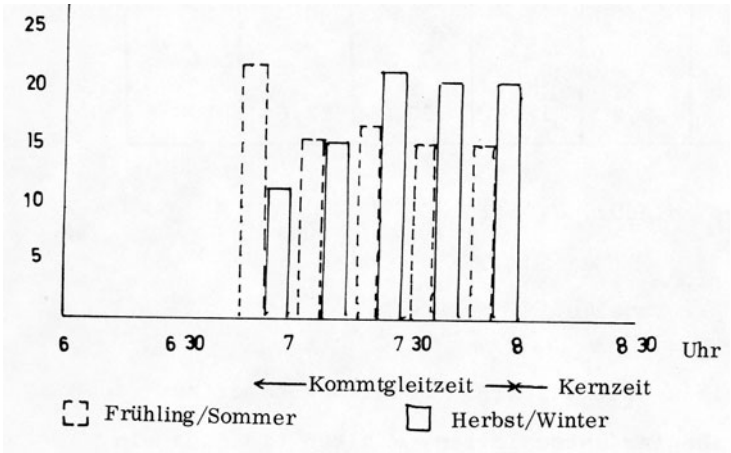


Abbildung 2

Anwesenheit des Personals (Betriebsbereitschaft)
in Prozent

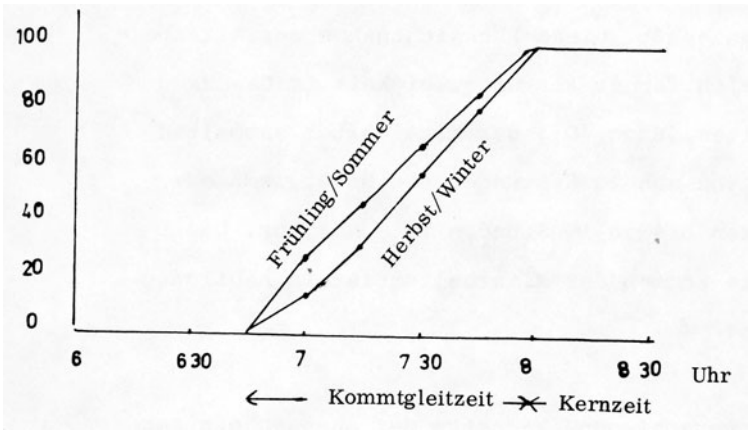


Abbildung 3

II. Fallstudie: Gleitzeitregelung bei stoßweisem Arbeitsanfall

A. Problemstellung

Der Personalrat eines Großstadtpostamtes erwägt die Möglichkeiten zur Einführung der gleitenden Arbeitszeit für die Telegramm- und Eilbriefboten, hauptsächlich, um die Arbeitsbedingungen dieser Gruppe attraktiver zu gestalten. Die Boten arbeiten in drei Schichten, wobei die Frühschicht, 6:00 - 14:00, mit 60 Personen am stärksten besetzt ist, die Nachtschicht {22:00 - 6:00} jedoch fast nur das für dringlichste Fälle benötigte Personal umfaßt.

Die Bedenken gegen eine Einführung der gleitenden Arbeitszeit richten sich lediglich auf Eigentümlichkeiten im Arbeitsanfall während der Frühschicht, siehe Tabelle 2. Insbesondere erscheint fraglich, ob bei einer gleitenden Arbeitszeit der während der Nacht eingetroffene Telegramm- und Eilbriefstapel, im Mittel 600 auszutragende Einheiten, ohne Stockungen abgebaut werden kann. Dieser Übertrag von der Nacht muß bis 8 Uhr das Haus verlassen haben. Dem entspricht eine Strafkostenfunktion für Verzögerungen von Nacht-eingängen, die keine Kosten vorsieht bei Zustellung vor 8 Uhr, und unendliche Kosten für spätere Zustellung. Würde die Mehrheit der Boten einen späteren Arbeitsbeginn als 6 Uhr wählen, so scheint ein Verzug unvermeidlich. Andererseits verspricht die gleitende Arbeitszeit eine größere Flexibilität im morgend-

lichen Abholen, denn bei der jetzigen Regelung, daß alle Boten um 6 Uhr erscheinen müssen, "treten sich die Leute ja auf die Füße".

Ankünfte von Telegrammen und Eilbriefen während der Frühschicht			
6 Uhr	600	= Übertrag v.d. Nacht	
6:00 - 6:30	30	10:00 - 10:30	170
6:30 - 7:00	38	10:30 - 11:00	125
7:00 - 7:30	54	11:00 - 11:30	100
7:30 - 8:00	70	11:30 - 12:00	120
8:00 - 8:30	115	12:00 - 12:30	160
8:30 - 9:00	170	12:30 - 13:00	190
9:00 - 9:30	240	13:00 - 13:30	165
9:30 - 10:00	240	13:30 - 14:00	135
	957		1155

Tabelle 2

Während der Frühschicht beträgt der durchschnittliche Anfall an Telegrammen und Eilbriefen

264/Stunde (ausschließlich Nachtübertrag) bzw.
339/Stunde (einschließlich Nachtübertrag).

Rechnet man nur die Zeit nach 8:00 Uhr, wenn die von der Nacht liegengebliebenen Einheiten ausgetragen sind, so beträgt der durchschnittliche Anfall pro Stunde 302 Telegramme und Eilbriefe. Nach 8 Uhr verläuft der Arbeitsanfall während der Frühschicht wie eine gedämpfte Schwingung um 300 Einheiten. Wegen dieser relativen Gleichmäßigkeit des Arbeitsanfalls

in den Stunden vor und nach 14 Uhr bereitet der Schichtwechsel um 14 Uhr auch bei der Einführung der gleitenden Arbeitszeit keine Schwierigkeiten. Hier überlappt sich die Gehrgleitzeit der Frühschicht mit der Kommtgleitzeit der Spätschicht, jedoch wird erwartet, daß dabei ein Ausgleich zwischen Kommen und Gehen eintritt, der die Betriebsbereitschaft aufrechterhält.

Die Einführung der gleitenden Arbeitszeit könnte also nur dann scheitern, wenn es nicht gelingt, den Anfangsbestand von 600 Einheiten während der ersten beiden Frühschichtstunden, d.h. bis 8 Uhr, abzubauen. Bei der bisherigen Arbeitszeitregelung, nach der alle Boten um 6 Uhr anwesend sein müssen, konnte der Nachtübertrag bei voller Betriebsbereitschaft reibungslos verteilt werden, weil der laufende Arbeitsanfall der ersten beiden Stunden sehr gering ist, siehe Tabelle 2. Läßt sich der Anfangsbestand auch bei verminderter Betriebsbereitschaft in den ersten Morgenstunden (6:00 - 8:00) bis 8 Uhr abbauen?

Speziell lautet die Frage, die vor einer Einführung einer morgendlichen Kommtgleitzeit befriedigend beantwortet sein muß:

1. Welche Lage und Dauer der Kommtgleitzeit gestattet den Abbau des Anfangsbestandes von 600 Telegrammen und Eilbriefen bis 8 Uhr?

Falls die Einführung einer Kommtgleitzeit die anfängliche Betriebsbereitschaft so mindert, daß das Verteilen des Übertrages bis 8 Uhr nicht gewährleistet ist, schließt sich folgende Frage an:

2. Welche Anzahl zusätzlicher Boten ist notwendig, um diesen Abbau bis 8 Uhr störungsfrei zu ermöglichen?

B. Alternative Gleitzeitmodelle

Um die Dispositionsvorteile der Mitarbeiter an der neuen Regelung zu maximieren, wäre eine möglichst lange Kommtgleitzeit erwünscht. Die Mindestlänge der Gleitspanne beträgt eine Stunde. Es werden drei Modelle der Gleitzeitregelung diskutiert:

Modell A: 60 Minuten Kommtgleitzeit

Modell B: 75 " "

Modell C: 90 " "

Neben der Dauer der Kommtgleitzeit ist auch ihre Lage zu bestimmen. Soll die Kommtgleitzeit um 5:30, 5:45 oder um 6 Uhr beginnen? 5:30 ist der früheste Beginn, da ein Austragen vor 6 Uhr nicht vorgesehen ist.

Wir wollen mit T die Kommtgleitzeit bezeichnen;
 und T_j bezeichnet die Kommtgleitzeiten der drei
 Modelle $j \in \{A, B, C\}$, nämlich

$$\begin{aligned} T_A &= \{5:30 - 6:30\} \\ T_B &= \{5:30 - 6:45\} \\ T_C &= \{5:30 - 7:00\} . \end{aligned}$$

Dabei ist zunächst 5:30 Uhr als Beginn der Gleitzeit
 angesetzt. Um auch noch nach den Anfangszeiten
 $k \in \{5:30, 5:45, 6:00\}$ aufschlüsseln zu können, sei die
 Kommtgleitzeit mit T_{jk} bezeichnet.
 Bei dieser Formulierung haben wir also die Vor- und
 Nachteile von neun Modellen der Gleitzeitregelung
 aufzuzeigen.

C. Das Bewertungsverfahren

Es kommt darauf an, für die einzelnen Modelle durchzu-
 spielen, inwieweit sie den Abbau des Anfangsbestandes
 bis 8 Uhr gewährleisten.

Dabei könnte man zunächst daran denken, ein Warte-
 schlangenmodell zur Analyse der Betriebsbereitschaft
 (= Abfertigung) und des Arbeitsanfalls (= Ankünfte im
 System) heranzuziehen. Unter der Annahme, daß es sich
 beim Eintreffen der Telegramme und Eilbriefe während
 der Frühschicht um einen stabilen Prozeß handelt,
 könnte man die durchschnittliche Ankunft von ca. 300
 Einheiten pro Stunde durch eine mittlere Ankunftsrate
 $\lambda = 5$ Einheiten pro Minute beschreiben. Das Zustellen

$\mu = 5$ Einheiten pro Minute (60 Telegrammboten erledigen 300 Einheiten pro Stunde).

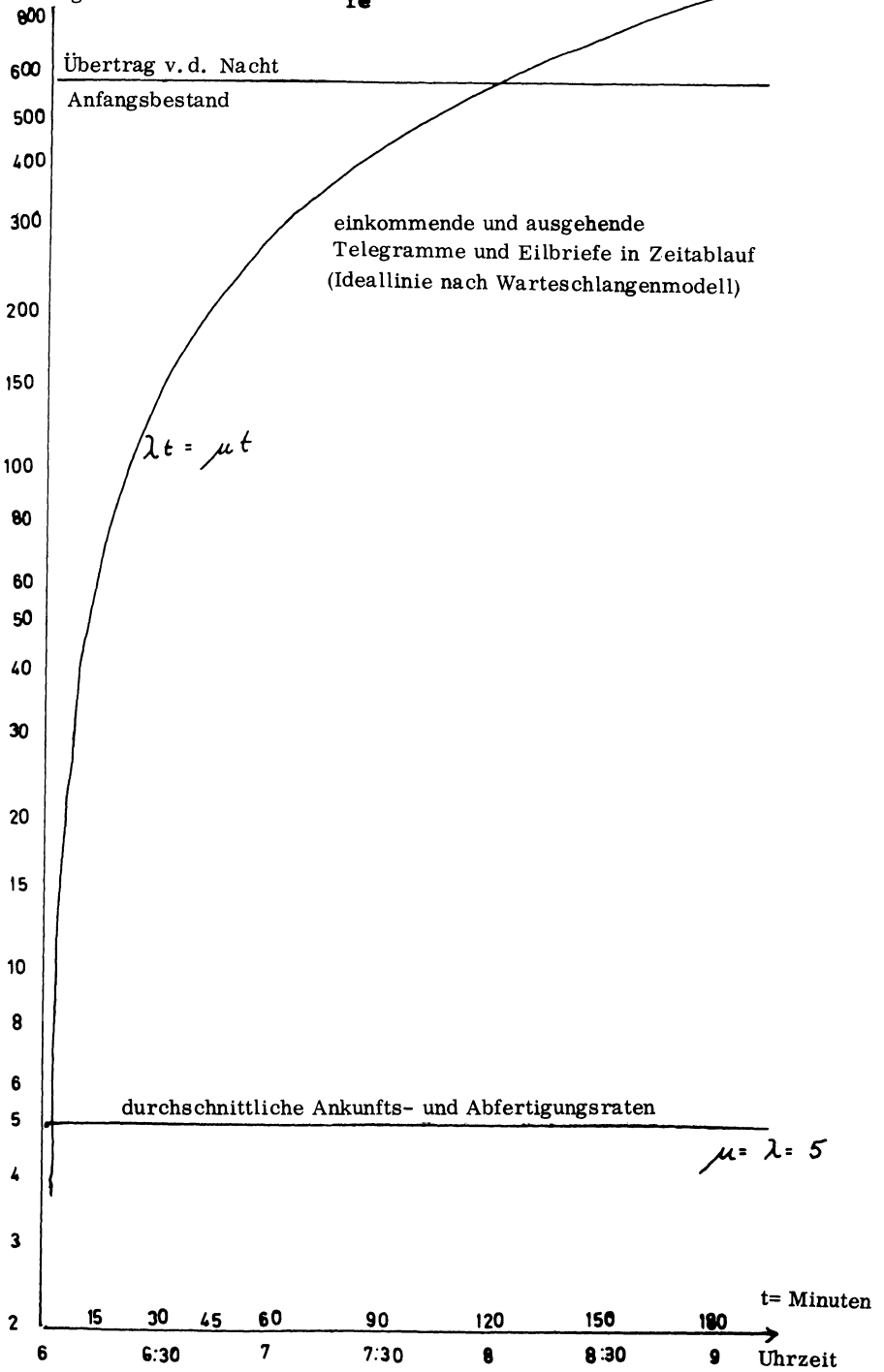
Bei der bestehenden Regelung der Arbeitszeit, Arbeitsbeginn 6:00 Uhr, Schluß 14:00 Uhr, wird der durchschnittliche Arbeitsanfall und die durchschnittliche Abfertigung durch λ und μ beschrieben, und der gesamte Arbeitsanfall und die gesamte Abfertigung bemißt sich nach λt bzw. μt , wobei $t = 1, 2, \dots, 480$ die fortlaufenden Minuten sind im Zeitabschnitt $\{6:00 - 14:00\}$.

Dieses stark idealisierte Verhalten des Systems verdeutlicht Abbildung 4. Dabei sind die beiden identischen Geraden λt und μt logarithmisch aufgetragen, d. h. gekrümmt abgebildet. Es zeigt sich, daß die Kapazität der Boten gerade ausreicht, um den Anfangsbestand von 600 Einheiten bis 8 Uhr abzubauen.

Der effektive Arbeitsanfall und die effektive Abfertigung gestaltet sich nicht so ideal. Der Arbeitsanfall schwankt über den Zeitraum $\{6:00 - 14:00\}$. Wie in Tabelle 2 gezeigt wird, wechseln Stoßzeiten und Flauten einander ab. Durch eine eindeutige Ankunftsrate λ kann diese Schwankung nicht befriedigend ausgedrückt werden, insbesondere nicht am Anfang der Frühschicht. Deshalb muß die Ankunftsrate als Funktion der Zeit angesetzt werden, z. B. als Treppenfunktion λ_t , mit t dem Zeitindex.

Anzahl Telegramme und Eilbriefe

fe



λ_t ergibt sich aus Tabelle 2. Punkt 6 Uhr fallen mit einem Schlage 600 Einheiten an; d. h., die Ankunftsrate λ_t hat hier, bei $t = 0$, ein erstes Maximum: $\lambda_0 = 600$. Dann entwickelt sich die Ankunftsrate λ_t von $\lambda_1 = 1$ zwischen 6:00 und 6:30 über $\lambda_{90} = \lambda_t = 2$ um etwa 7:30 auf $\lambda_{120} = \lambda_t = 3$ um etwa 8 Uhr. Sie erreicht ein zweites Maximum um etwa 9:30 mit $\lambda_{210} = \lambda_t = 8$, siehe Abbildung 5. Im hier interessanten Zeitabschnitt, zu Anfang der Frühschicht, ist der Ankunftsprozeß keinesfalls stationär, obwohl er sich später auf die durchschnittliche Ankunftsrate λ einpendelt. Gleichfalls in Abbildung 5 ist die effektive Abfertigung eingezeichnet, wie sie sich bei gleitender Arbeitszeit gestalten würde. Dabei sind die Abfertigungsraten wiederum von der Zeit abhängig, und zwar nach Modellen $j \in \{A, B, C\}$ verschieden. Je nachdem, ob eine Kommtgleitzeit T_A , T_B , oder T_C angenommen wird, entwickelt sich die Betriebsbereitschaft verschieden. Deshalb gilt zu verschiedenen Zeiten auch eine unterschiedliche Abfertigungsrate μ_t .

		5:30- 5:45	5:45- 6:00	6:00- 6:15	6:15- 6:30	6:30 6:45	insge- samt
1	Eintreffen, in % der Mitarbeiter	12,5	15	22,5	25	25	100
2	Eintreffen, in Teilen von $\mu = 5$	0,625	0,75	1,125	1,25	1,25	5
3	Anwesenheit, in akkumulier- ten Teilen von $\mu = 5$	0,625	1,375	2,5	3,75	5	5

Tabelle 3

Dies sei am Modell B, Kommtgleitzeit T_B von 5:30 bis 6:45 Uhr, illustriert. Wir nehmen zunächst an, daß die Mitarbeiter ein ähnliches Kommtverhalten entwickeln werden wie es für andere Organisationen typisch ist, und vermuten, daß das Kommtverhalten der Mitarbeiter des Landratsamtes Forchheim repräsentativ ist, siehe Tabelle 1.

Das Kommtverhalten dieser Gruppe zeigte zwei saisonabhängige Profile: ein früheres Anwachsen der Betriebsbereitschaft im Frühjahr und Sommer, und ein späteres Kommen in Herbst und Winter. Um Risiken abzuschwächen, wählen wir das Profil mit verzögertem Kommtverhalten als Vorbild. Wenn bei pessimistischem Ansatz des Anwachsens der Betriebsbereitschaft keine Störungen auftreten, dann sollte eigentlich die Wahrscheinlichkeit für Störungen in Wirklichkeit geringer sein.

Um das Risiko weiter einzugrenzen, schlagen wir den Anteil von 8 % der täglich wechselnd kommenden Mitarbeiter der Periode (7:30 bis 8:00) zu (Tabelle 1) und teilen diese in zwei Unterperioden auf, siehe Tabelle 3, Zeile 1. In Zeile 2 werden die Prozentangaben der Zeile 1 als analoge Teile von $\mu = 5$ ausgedrückt; in Zeile 3 werden diese Teile akkumuliert. Die Zeiteinteilung in fünf Viertelstunden, der Länge von T_B , erlaubt die Datierung des vermuteten Kommtverhaltens bzw. des Aufbaus der Betriebsbereitschaft gemessen als Anwachsen der Abfertigungsrate zwischen 0 und 5.

Die Ziffern der Tabelle 3, Zeile 3, bewerten die verschiedenen Stadien im Aufbau der Betriebsbereitschaft; sie sind Maße für die Fähigkeit zum Austragen von Telegrammen und Eilbriefen in den Intervallen $T_{B1} = \{5:30 - 5:45\}$, $T_{B2} = \{5:46 - 6:00\}$, ..., $T_{B5} = \{6:31 - 6:45\}$.

Die Abfertigungsraten in den jeweiligen Intervallen nennen wir μ_{ji} , $i = 1, 2, \dots, 5$, $j = B$, und diese Werte $\mu_{B1} = 0,625$, $\mu_{B2} = 1,375$, ..., $\mu_{B5} = 5$ sind als Kurve μ_B in Abbildung 5 eingezeichnet. Damit ist, für das Modell B, das Kapazitätswachstum während der Kommtgleitzeit beschrieben.

Ab 6:45, dem Ende der Gleitzeit und Anfang der Kernzeit (volle Anwesenheit), gilt $\mu_t = \mu = 5$ für $t \geq 75$, d.h. $t \in \{6:45 - 14:00\}$.

Entsprechend ist, für Modell A und C:

$$\begin{aligned} \mu_{A1} &= 0,625 \quad \text{im Intervall } T_{A1} = \{5:30 - 5:41\} \\ \mu_{A2} &= 1,375 \quad " \quad " \quad T_{A2} = \{5:42 - 5:53\} \end{aligned}$$

⋮

$$\begin{aligned} \mu_{A5} &= 5 \quad " \quad " \quad T_{A5} = \{6:31 - 6:30\} \\ \text{und} \end{aligned}$$

$$\begin{aligned} \mu_{C1} &= 0,625 \quad \text{im Intervall } T_{C1} = \{5:30 - 5:47\} \\ \mu_{C2} &= 1,375 \quad " \quad " \quad T_{C2} = \{5:48 - 6:05\} \end{aligned}$$

⋮

$$\mu_{C5} = 5 \quad " \quad " \quad T_{C5} = \{6:42 - 7:00\}.$$

Mit anderen Worten, je nachdem, ob wir bei einem

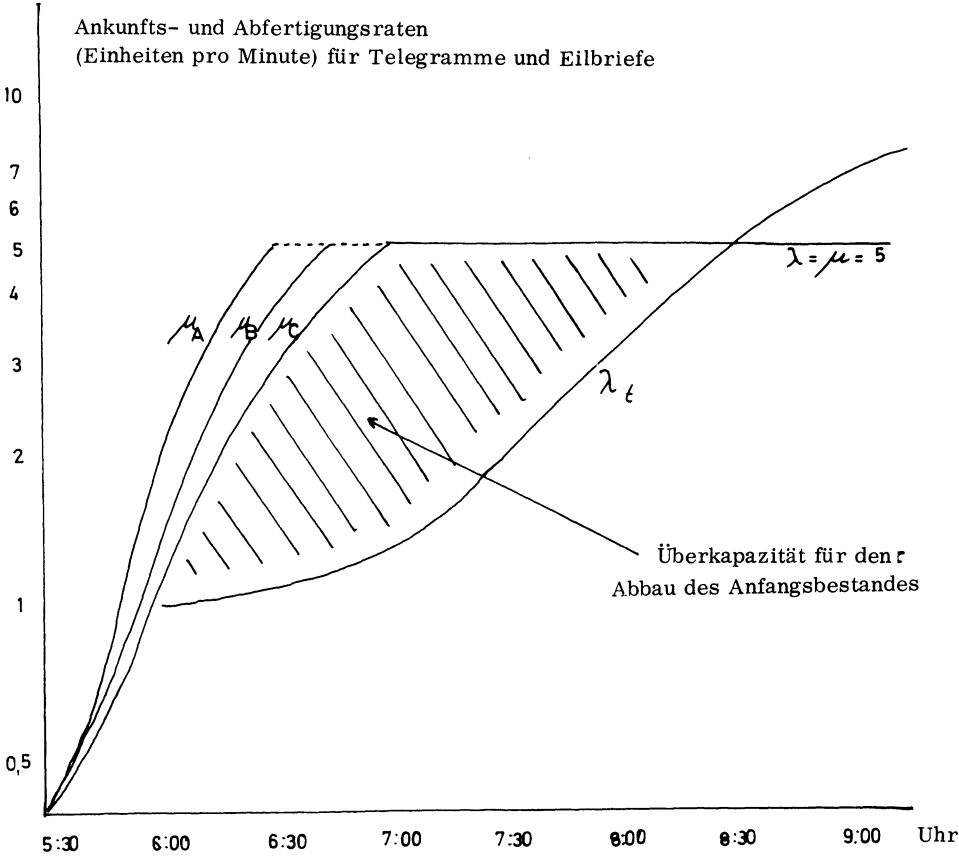


Abbildung 5

Beginn um 5:30 eine Gleitzeit von einer Stunde, von eineinviertel oder von eineinhalb Stunden (Modelle A, B, C) ansetzen, ergibt sich in Abbildung 5 eine unterschiedliche Abfertigungsrate (Kurven μ_A, μ_B, μ_C). Dabei sind die Raten μ_{j1} innerhalb der Teilabschnitte T_{j1} eingezeichnet, d. h. im Abstand von d_j , $j \in (A, B, C)$, wobei

$$d_A = 12 \text{ Minuten}$$

$$d_B = 15 \text{ Minuten}$$

$$d_C = 18 \text{ Minuten.}$$

Aus dem gemeinsamen Bild 5 der Ankunfts- und Abfertigungsraten wird das Entscheidungsproblem noch einmal deutlich. Bei der jetzigen, starren Arbeitszeitregelung, Beginn 6 Uhr, Schluß 14 Uhr, gelingt es dem Personal, während der zwei Arbeitsstunden bis 8 Uhr den Anfangsbestand von 600 Einheiten abzubauen. Dafür ist die Überkapazität, die im Bild schraffierte Lücke zwischen Betriebsbereitschaft und effektivem Arbeitsanfall (ohne Anfangsbestand) gerade ausreichend. Wird durch Einführen der gleitenden Arbeitszeit diese Lücke nicht zu sehr beschnitten?

Zur Beantwortung muß die gesamte Abfertigung und der gesamte Arbeitsanfall verglichen werden.

So wie die gesamte Abfertigung im Idealfall gemäß μ_t verläuft, entwickelt sich die effektive Gesamtabfertigung nach Maßgabe der μ_{j1} . Die gesamte Abfertigung $S_A(t)$ bei Modell A ist als Kurve A in

Abbildung 6 dargestellt. Sie entsteht durch
Akkumulation wie folgt:

$$\begin{aligned}
 S_A(t) &= \mu_{A1} t && \text{für } t \in T_{A1} \\
 &&& \text{d. h. } t = 1, 2, \dots, 12 \\
 S_A(t) &= 12 \mu_{A1} + \mu_{A2} t && \text{für } t \in T_{A2} \\
 &&& \text{d. h. } t = 13, 14, \dots, 24 \\
 &\vdots && \\
 S_A(t) &= 12 \sum_{i=1}^4 \mu_{Ai} + \mu_{A5} t && \text{für } t \in T_{A5} \\
 &&& \text{d. h. } t = 49, 50, \dots, 60
 \end{aligned}$$

Entsprechend berechnet man die Kurven $j = B, C$ für
die Modelle B und C:

$$S_j(t) = d_j \sum_{i=1}^{u-1} \mu_{ji} + \mu_{jn} t \quad \text{für } t \in T_j$$

Während der Kernzeit entwickelt sich die gesamte
Abfertigung gemäß μt fort, siehe Abbildung 6.

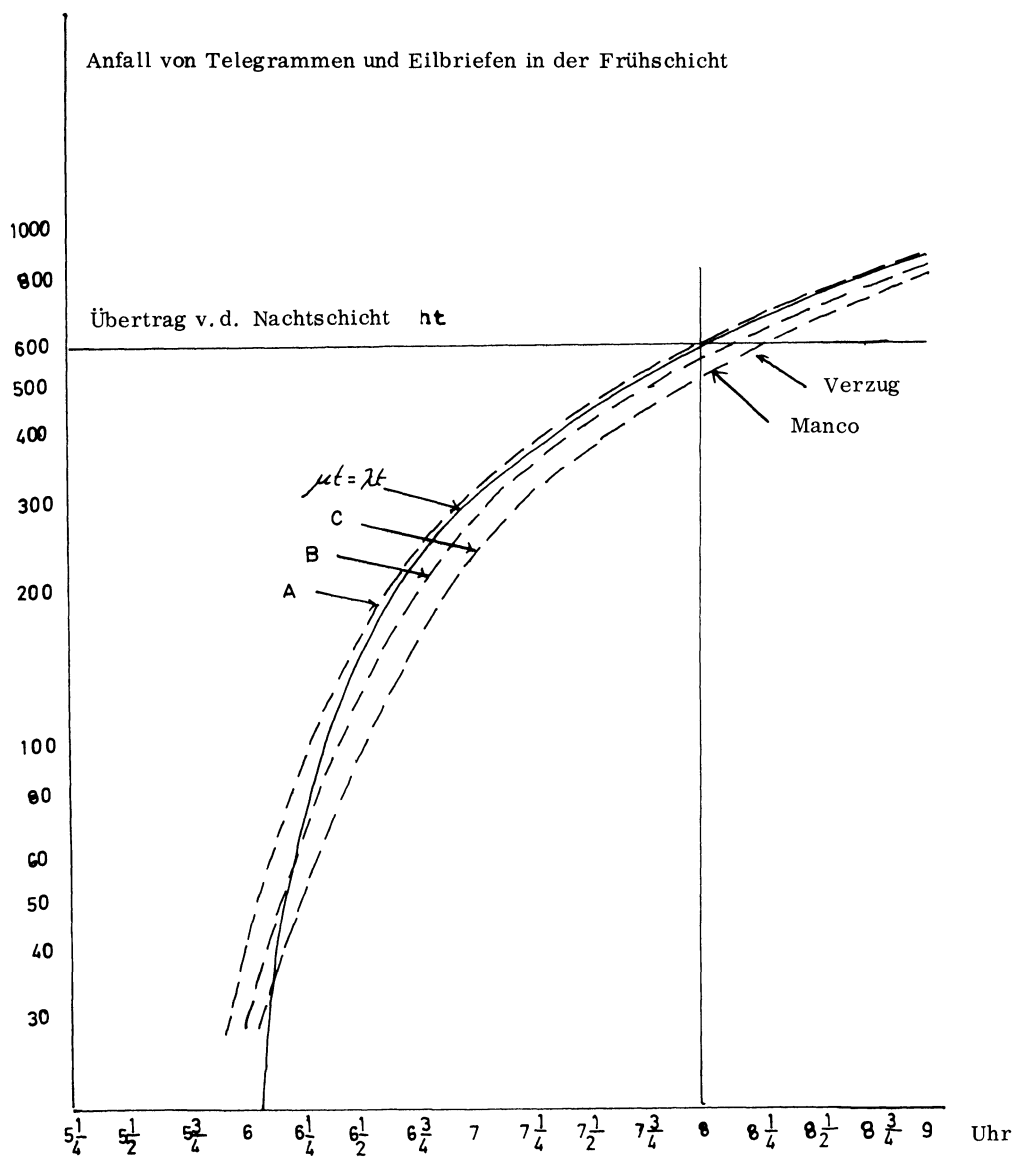


Abbildung 6

D. Ergebnisse

Die folgenden Aussagen können der Abbildung 6
direkt entnommen werden:

E 1: Die gleitende Arbeitszeit mit einer Kommtgleit-
====
zeit $T_A = \{5:30 - 6:30\}$ ist der gegenwärtigen
Regelung -Arbeitsbeginn für alle ist 6 Uhr-
überlegen.

Kurve A liegt stets über Kurve μ_t , d.h. der Anfangs-
bestand wird früher abgebaut.

Weil eine längere Kommtgleitzeit die Vorteile der
Mitarbeiter erhöht, fragen wir nun:

Läßt sich eine Verlängerung der Kommtgleitspanne
 $T_A = \{5:30 - 6:30\}$ um 15 oder 30 Minuten vertreten?

E 2: Eine Kommtgleitzeit $T_B = \{5:30 - 6:45\}$ führt zu
====
einer Verzögerung von etwa 8 Minuten.

Kurve B schneidet die Linie 600 bei ca. 8:08 Uhr;
um 8:00 Uhr sind noch etwa 26 Einheiten vorhanden
(4,33 % v. 600). Um die Verzögerung zu verhindern, müssen
2,6 Boten (4,33 % v. 60) neu eingestellt werden.

E 3: Eine Kommtgleitzeit $T_C = \{5:30 - 7:00\}$ führt zu
====
einer Verzögerung von ca. 16 Minuten.

Kurve C schneidet B = 600 bei ca. 8:16 Uhr; um 8:00
sind noch etwa 71 Einheiten vorhanden (11,83 % v. 600).

Um die Verzögerung zu verhindern, müssen 7,1 Boten (11,83 % v. 60) neu eingestellt werden. Aufgrund der gespannten Arbeitsmarktlage und der Betriebskostenstruktur kommt die Einstellung von 7 neuen Boten und die Lohnkostensteigerung von 12 % nicht infrage.

Untersuchen wir nun die Frage, ob eine Verlegung der morgendlichen Kommtgleitzeit von 5:30 auf später wünschenswert ist. Diese Verlegung entspricht - in Abbildung 6 - einer Parallelverschiebung der Kurven A, B, C.

E 4: Eine Kommtgleitzeit {5:45 - 6:45} führt zu einer
====
Verzögerung von etwa 15 Minuten. Die Verlegung der Kommtgleitzeit um 15 Minuten führt zu ähnlichen Verzögerungen wie das Verlängern um 30 Minuten (E 3). Deshalb kommt eine Kommtgleitzeit, die nicht schon um 5:30 Uhr beginnt, nicht in Betracht.

Das Entscheidungsproblem über die optimale Gleitzeitregelung reduziert sich - im vorliegenden Fall - mithin auf die beiden Möglichkeiten der Kommtgleitzeit:

{5:30 - 6:30} oder {5:30 - 6:45}?

Dabei müssen die zusätzlichen Dispositionsvorteile, die den Mitarbeitern durch die Verlängerung der Gleitzeit entstehen, mit den zusätzlichen Lohnkosten (ca. 4,33 %) abgewogen werden. Besteht keine Aussicht, daß infolge des Arbeitskräftemangels weitere Mitarbeiter gewonnen

werden können, dann muß die einstündige Morgen-Kommtzeit gewählt werden. Sie ist der gegenwärtigen Regelung mit fixen Arbeitszeiten überlegen.

Literatur

- 1 Hackh, S., Gleitende Arbeitszeit
Verlag moderne Industrie, München, 1971
- 2 Knevels, P., und R. Zehle, Variable Arbeitszeit,
Heider-Verlag, Bergisch Glädbach, 1971

Ein Simulationsmodell für den Instandhaltungsbereich

von D. Ordelheide, Bochum

1 Einführung in die Problemstellung

Das Interesse der Unternehmensforschung an Instandhaltungsproblemen hat sich in den letzten Jahren verstärkt. Man bemühte sich u.a. um folgende Entscheidungskomplexe:

- Sollen die Teile technischer Systeme erst ersetzt werden, wenn sie funktionsunfähig sind oder soll man sie vorbeugend auswechseln? Welche der möglichen Vorbeugungsstrategien soll gewählt werden?
- Wie groß soll man die Kapazitäten der Instandhaltungsabteilungen machen?
- In welcher Reihenfolge sollen Instandhaltungsaufträge, die vor belegten Instandhaltungskapazitäten warten, abgefertigt werden?

Das nachfolgende Simulationsmodell soll dazu dienen, diese und weitere Instandhaltungsentscheidungen untersuchen zu können. Es werden damit die Bemühungen von Vergin ¹⁾, Frotscher ²⁾ und Kress ³⁾ fortgesetzt.

+) Institut für Unternehmensführung und Unternehmensforschung der Ruhr-Universität Bochum, 4630 Bochum-Querenburg, Buscheystraße

++) Überarbeitetes Manuskript eines auf der DGU-Tagung vom 22. - 24. 9. 1971 gehaltenen Vortrages.

2 Modellelemente

2.1 Produktionsanlagen

Der Produktionsbereich des Modells kann aus einer oder mehreren Fertigungsanlagen bestehen. Jede der Anlagen kann mehrere Reparaturteile aufweisen. Jedes Reparaturteil kann einen der beiden Zustände "funktionsfähig" und "nicht funktionsfähig" haben. Zwischenzustände (= verschiedene Grade der Funktionsfähigkeit) sind nicht möglich. Das Teil ist definiert durch

- eine Wahrscheinlichkeitsverteilung seiner Laufzeit. Laufzeit ist die Zeit(gemessen in Betriebs- oder Nutzungszeiteinheiten) zwischen dem Einbau des Teiles in die Produktionsanlage und dem Moment der Funktionsunfähigkeit.
- eine Wahrscheinlichkeitsverteilung der Reparaturzeit für die Schadensreparatur des Teiles. Schadensreparatur ist die Wiederherstellung der Anlage durch Ersatz des Teiles nach Eintritt seiner Funktionsunfähigkeit. Wird das Teil vorher ausgewechselt, spricht man von vorbeugender Reparatur. Das Ersatzteil gehört zur gleichen Grundgesamtheit wie das ursprüngliche Reparaturteil (sog. as-new-Annahme).
- eine konstante Dauer für die vorbeugende Reparatur
- eine Kennziffer der für die Reparatur des Teiles zuständigen Instandhaltungsabteilung
- die Anzahl der Arbeitskräfte dieser Abteilung, die für die Reparatur des Teiles benötigt werden und
- drei Reparaturstrategie-Variable (Erläuterung siehe Abschnitt 2.3).

Die Reparaturteile sind sogenannte kritische Teile, d. h. wird eines von ihnen funktionsunfähig oder wird es vorbeugend überholt oder

ausgetauscht, so steht während dieser Zeit die gesamte Anlage still; die Laufzeiten aller anderen Teile der Anlage sind unterbrochen (sog. ökonomische Abhängigkeit der Reparaturteile einer Anlage).

2.2 Instandhaltungsorganisation

Die Instandhaltungsorganisation besteht im Modell aus einer oder mehreren Instandhaltungsabteilungen. Die Abteilungsbildung innerhalb der Gesamtorganisation kann nach unterschiedlichen Kriterien erfolgen (z. B. Berufsausbildung und/oder Produktionsbetrieben und/oder Anlagentypen). Die Größe der Abteilung ist durch die Anzahl ihrer einsatzfähigen Handwerker bestimmt. Der einzelne Handwerker kann nicht identifiziert werden. Es kann jedoch innerhalb der Gesamtkapazität einer Abteilung zwischen Basiskapazität und Überstundenkapazität unterschieden werden. Die Handwerker der Überstundenkapazität werden erst dann beansprucht, wenn alle Handwerker der Basiskapazität beschäftigt sind. Diese Art der Bildung von Teilkapazitäten ist z. B. im Hinblick auf die Lohnkostenermittlung bedeutsam. Bezüglich aller anderen im Modell angesprochenen Eigenschaften sind die einzelnen Mitglieder einer Instandhaltungsabteilung identisch.

2.3 Reparaturstrategien

Eine Reparaturstrategie legt fest, zu welchen Zeitpunkten des Planungszeitraumes ein Reparaturteil ersetzt oder wiederhergestellt werden soll. Im Modell werden folgende Reparaturstrategien unterschieden:

Feuerwehrstrategie: Bei dieser Strategie werden die Reparaturteile immer erst nach Eintritt ihrer Funktionsunfähigkeit ersetzt.

t_i, α -Strategien: Für das i -te Reparaturteil ($i = 1, 2, \dots, n$) werden äquidistante Zeitpunkte $(0+t_i)$, $(0+2t_i)$, $(0+3t_i)$, ... festgelegt. Zu diesen sogenannten Vorbeugungszeitpunkten wird das Alter des Teiles (gemessen in Betriebszeiteinheiten) festgestellt. Ist ein bestimmtes Mindestalter erreicht, wird das Teil vorbeugend ersetzt. Das Mindestalter ist definiert als

$$\alpha_i (t_i - t_i^V) \quad 0 \leq \alpha_i \leq 1; \quad t_i^V < t_i < \infty.$$

t_i ist die Länge des Zeitintervalls (gemessen in Kalenderzeiteinheiten) zwischen zwei Vorbeugungszeitpunkten (= Vorbeugungsintervall). t_i^V ist die Dauer der vorbeugenden Überholung oder Ersetzung des i -ten Reparaturteiles ohne Wartezeit. $t_i - t_i^V$ ist dann der Teil des Vorbeugungsintervalls, der überhaupt maximal für die Nutzung des Teils zur Verfügung steht. Bei $\alpha_i = 1$ muß das Teil diesen gesamten Zeitraum benutzt worden sein, wenn zum Ende des Vorbeugungsintervalls eine vorbeugende Reparatur vorgenommen werden soll. Bei $\alpha_i = 0$ ist das Mindestalter gleich 0, so daß in jedem Vorbeugungszeitpunkt vorbeugend repariert wird, es sei denn es liefe für dieses Teil schon eine Schadensreparatur. Diese wird, sowie Reparaturkapazität frei ist, immer unmittelbar nach Eintritt der Funktionsunfähigkeit des Teils vorgenommen.

Für ein bestimmtes Reparaturteil i sind t_i und α_i Aktionsvariable, t_i^V ist eine Konstante. Für $\alpha = 0$ sind die t, α -Strategien der aus der Erneuerungstheorie bekannten Klasse der streng periodischen Strategien vergleichbar. Allerdings fehlt dort die Verknüpfung des erneuerungstheoretischen mit dem warteschlangentheoretischen Problem (vgl. Abschnitt 3). Für $\alpha = 1$ ähneln die t, α -Strategien den bekannten altersabhängigen Vorbeugungsstrategien. Wählt man das Mindestalter größer als die maximale Laufzeit des Reparaturteils (Obergrenze der Laufzeitverteilung), so entspricht die t, α -Strategie der Feuerwehrstrategie. Existiert keine Laufzeitobergrenze, müßte man es größer als die Planungsperiode wählen. Diese allgemeinen Eigenschaften der Klasse der t, α -Strategien machen sie für Simulationsstudien besonders geeignet, da sich nur durch unterschiedliche Wertzuweisung an die Parameter α_i und t_i unterschiedliche Klassen von Strategien abbilden lassen.

Blockstrategien: Wählt man für zwei oder mehrere Reparaturteile gleich lange Vorbeugungsintervalle und beginnen die ersten Vorbeugungsintervalle dieser Teile gleichzeitig, so fallen für den gesamten Planungszeitraum die Vorbeugungszeitpunkte dieser Teile zusammen. Durch Blockbildungen dieser Art kann man möglicherweise instandhaltungsbedingte Produktionsausfälle herabsetzen.

Andererseits besteht die Gefahr, daß durch Festlegung der Vorbeugungsintervalle unerwünschte Reparaturhäufungen in bestimmten Zeitpunkten auftreten,

so daß die zur Durchführung aller dieser Reparaturen notwendige Kapazität größer als die Maximalkapazität bestimmter Instandhaltungsabteilungen ist. Dies kann selbst für Teile mit gleichlangen Vorbeugungsintervallen dadurch vermieden werden, daß man die Vorbeugungszeitpunkte staffelt. Dies wird im Modell durch das Instrument des Anfangsintervalls ermöglicht. Das Anfangsintervall eines Reparaturteils ist der Zeitraum zwischen dem Beginn der Simulationszeit und dem ersten Vorbeugungszeitpunkt für dieses Reparaturteil.

Beispiel: Es sei von zwei Reparaturteilen ausgegangen. Das Vorbeugungsintervall betrage für beide fünf Zeiteinheiten (=ZE). Eine vorbeugende Reparatur für jedes Teil beansprucht zwei Handwerker derselben Kapazität eine ZE lang, Insgesamt stehen maximal drei Handwerker zur Verfügung. Wählt man die Anfangsintervalle für beide Teile gleich lang, so muß bei Erreichen des Vorbeugungszeitpunktes und falls beide Teile das Mindestalter überschritten haben, eines der beiden Teile eine ZE lang auf seine Reparatur warten. Legt man das Anfangsintervall für das eine Teil z. B. mit drei und für das andere Teil mit fünf ZE fest, so entfällt diese Wartezeit.

Zusätzlich zu den in Abschnitt 2.1 im einzelnen aufgeführten Merkmalen wird ein Reparaturteil somit durch die Strategievariablen a_i , t_i und das Anfangsintervall definiert.

2.4 Prioritätsregeln

Prioritätsregeln legen die Reihenfolge fest, in der Instandhaltungsaufträge, die vor belegten Instandhaltungskapazitäten warten, abgefertigt werden. Da in diesem Modell - im Unterschied zu den bekannten Warteschlangenmodellen - verschiedene Reparaturaufträge die Instandhaltungskapazitäten auch in unterschiedlichem Ausmaß beanspruchen können (s. Abschnitt 2.1), so konkurrieren bei Freisetzung von Kapazitätseinheiten einer bestimmten Instandhaltungsabteilung nicht notwendig alle vor dieser Abteilung wartenden Instandhaltungsaufträge um die freie Kapazität. Ist z. B. eine Reparatur beendet, für die 2 Elektriker benötigt wurden, und sind sonst keine Elektriker verfügbar, so zählen zu den konkurrierenden Aufträgen lediglich jene, die einen oder zwei Elektriker erfordern. Die Prioritätsregeln werden im folgenden für diesen allgemeineren Fall definiert.

Als Prioritätsregel wurde im vorliegenden Modell die SPT-Regel (shortest-processing-time) gewählt. Danach wird immer der Auftrag mit der kürzesten Reparaturzeit zuerst bearbeitet. Bei vorbeugenden Reparaturen ist die Zeitdauer bekannt; bei Schadensreparaturen wird die erwartete Reparaturzeit genommen.

2.5 Instandhaltungsziele

Das Verhalten des Systems bei verschiedenen Reparaturstrategien wird an folgenden Zielgrößen gemessen:

- Stillstandsrate der Produktionsanlagen. Sie ist gleich dem Verhältnis der instandhaltungsbedingten Stillstandszeit (= instandhaltungsbedingte Warte- und Reparaturzeit) zur maximal möglichen Produktionszeit einer Anlage in der Planungsperiode.
- Auslastungsgrad der Reparaturabteilungen. Er ist gleich dem Quotienten aus der Menge der tatsächlich von allen Handwerkern einer Abteilung geleisteten Arbeitsstunden und der Gesamtzahl der Arbeitsstunden, die sie maximal hätten eingesetzt werden können. Eine Zerlegung in einen Auslastungsgrad der Basiskapazität und einen Auslastungsgrad der Überstundenkapazität ist für jede Abteilung möglich (s. Abschnitt 2.2).
- Gesamtverbrauch an Ersatzteilen.

3 Modellablauf

Den Modellablauf kann man grob wie folgt beschreiben:

Definitionsteil: Zu Beginn eines Simulationslaufs oder einer Serie von Läufen werden die das Produktionssystem, den Instandhaltungsbetrieb und die Strategievariablen definierenden Merkmale numerisch festgelegt. Sie bestimmen zusammen mit den Zufallszahlen-Generatoren zur Erzeugung der Lauf-

zeiten und der Zeiten für Schadensreparaturen den weiteren Modellablauf. Das Gesamtsystem befindet sich bisher noch im Simulationszeitpunkt 0.

Produktionsbereich: Im Produktionsbereich des Modells wird für jede Anlage die Produktion abgebildet. Sie äußert sich in einem Verstreichen der Laufzeiten der Reparaturteile der Anlage. Die erste Produktionsphase beginnt im Simulationszeitpunkt 0 und endet mit dem ersten Ausfall oder mit dem Erreichen des ersten Vorbeugungszeitpunktes der Produktionsanlage. Der erste Ausfallzeitpunkt und der erste Vorbeugungszeitpunkt der Produktionsanlage sind durch die kürzeste Laufzeit bzw. das kürzeste Anfangsintervall aller Reparaturteile der Anlage festgelegt. Ist die kürzeste Laufzeit aller Reparaturteile der Anfangsgeneration niedriger als das kürzeste Anfangsintervall aller Teile, so wird die erste Produktionsphase der Anlage durch einen Ausfall beendet, dem dann eine Schadensreparatur folgt; andernfalls wird die Anlage erstmals zum Zwecke einer vorbeugenden Reparatur stillgelegt. Die Simulationszeit beträgt am Ende der ersten Produktionsphase $(0+t_p)$ wobei t_p gleich dem Minimum über alle Laufzeiten der ersten Generation und den Anfangsintervallen aller Reparaturteile ist.

Reparaturbereich: Im Reparaturbereich des Modells wird das Warten einer Anlage auf ihrer Reparatur und deren Durchführung abgebildet. Zu Beginn wird geprüft, ob eine ausreichende Anzahl an Arbeitskräften der zuständigen Instandhaltungsabteilung frei ist. Ist das der Fall, so wird die Anlage umgehend repariert. Die Reparatur äußert sich in einem Verstreichen der Reparaturzeit des auszuwechselnden Reparaturteiles. Bei Schadensreparaturen handelt es sich dabei um eine mit Hilfe der Reparaturzeitverteilung erzeugte Reparaturzeit. Bei vorbeugenden Reparaturen wird die geplante konstante Reparaturdauer (s. Abschnitt 2.1) des Reparaturteils genommen. Nach Beendigung der Reparatur befindet sich das System im Simulationszeitpunkt $(0+t_p + t_r)$, wobei t_r gleich der Reparaturzeit ist. Steht keine ausreichende Anzahl an Reparaturkräften zur Verfügung, so müssen das Teil und damit die gesamte Produktionsanlage auf die Reparatur warten. Ist die Reparatur für ein anderes Teil beendet, so belegen die wartenden Reparaturteile in Abhängigkeit von der gewählten Prioritätsregel die freiwerdende Kapazität.

Bevor der Ersatz eines ausgefallenen Teiles beendet ist, können vorbeugende Reparaturen (nicht Schadensreparaturen) für andere Teile der Anlage anstehen. Sie durchlaufen dann ebenfalls in der dargestellten Art und Weise den Instandhaltungsbereich.

Am Ende der Reparaturphase erhalten alle überholten Reparaturteile der Anlage eine neue Laufzeit und gegebenenfalls wird ihr nächster Vorbeugungszeitpunkt neu berechnet. Die Anlage ist dann wieder produktionsbereit und ein neuer Produktions-Instandhaltungs-Zyklus kann beginnen.

Demonstrationsbeispiel:

Der Produktionsbereich bestehe aus zwei Anlagen. Die beiden Anlagen sollen je zwei Reparaturteile aufweisen. Die Laufzeiten für die ersten zwei Generationen der Reparaturteile, die ersten zwei Reparaturzeiten für Schadensreparaturen, die Dauern der vorbeugenden Reparaturen, die zuständigen Reparaturabteilungen, die Werte für die Anzahl der benötigten Arbeitskräfte, die α -Werte, die Vorbeugungsintervalle und die Anfangsintervalle sind in Tabelle 1 angegeben. Sie definieren das Produktionssystem. Der Instandhaltungsbereich bestehe aus zwei Reparaturabteilungen, die erste mit zwei und die zweite mit drei Arbeitskräften.

Tab. 1: Definition eines Produktionssystems

Anlage	1		2	
Reparaturteil	1	2	3	4
Laufzeiten (simuliert)	25, 23	12, 17	22, 28	33, 15
Zeiten für Schadensreparaturen (simuliert)	3, 2	2, 3	6, 5	2, 4
Dauer vorbeugende Reparatur	1	2	3	1
Zuständige Reparaturabteilung	1	2	1	2
Zahl der benötigten Kräfte	2	1	2	3
α	0,5	0,5	0,5	0,5
Vorbeugungsintervall	24	28	30	28
Anfangsintervall	24	15	30	27

Der Modellablauf ist in Tab. 2 dargestellt:

Die erste Zeile enthält die Simulationszeit. Die Zeilen 2 bis 8 bestimmte Zeitgrößen für die Anlage 1, die Zeilen 9 bis 15 die entsprechenden Größen für die Anlage 2 und die Zeilen 16 und 17 geben die Belegung der Instandhaltungskapazitäten an.

Im Simulationszeitpunkt 0 entsprechen die kumulierten Laufzeiten der vier Reparaturteile ihren in Tab. 1 angegebenen ersten Laufzeiten; die kumulierten Vorbeugungsintervalle sind gleich den in Tab. 1 angegebenen Anfangsintervallen. Als Restlaufzeit eines Reparaturteils sei die Differenz zwischen seiner kumulierten Laufzeit und der bisherigen Laufzeit der Anlage definiert. Das Restvorbeugungsintervall eines Reparaturteils ist gleich dem kumulierten Vorbeugungsintervall abzüglich der gegenwärtigen Simulationszeit. Vergleicht man für die Anlage 1 die kleinste Restlaufzeit ($12 - 0 = 12$) mit dem kleinsten Restvorbeugungsintervall ($15 - 0 = 15$) so erhält man als erstes Ereignis den Ausfall von T_2 im Zeitpunkt 12. Für Anlage 2 ergibt sich entsprechend als erstes Ereignis der Ausfall von T_3 im Zeitpunkt 22.

Im Zeitpunkt 12 fällt T_2 aus und wird von einem Handwerker der Instandhaltungsabteilung 2. repariert. Die bisherige Laufzeit der Anlage 1 erhält den Wert 12 (Zeile 6, Spalte 2).

Im Zeitpunkt 14 ist die Reparatur von T_2 beendet. Die kumulierte Laufzeit wird neu berechnet, indem zur bisherigen kumulierten Laufzeit die mit Hilfe der Laufzeitverteilung erzeugte neue Laufzeit der zweiten Generation von T_2 im Wert von 17 (s. Tabelle 1) hinzugefügt wird. Es ergibt sich ein Wert von 29 (Zeile 3, Spalte 3). Der Wert des kumulierten Vorbeugungsintervalls braucht nicht verändert zu werden, da er noch größer als die gegenwärtige Simulationszeit ist. Das nächste Ereignis für die Anlage 1 ergibt sich aus einem Vergleich der kleinsten Restlaufzeit ($T_1: 25 - 12 = 13$) mit dem kleinsten Restvorbeugungsintervall ($T_2: 15 - 14 = 1$). Die Anlage geht somit bis zum Erreichen des ersten Vorbeugungszeitpunktes für T_2 (15) in Produktion.

Tab. 2 Simulation eines Produktions- und Instandhaltungsprozesses

		1	2	3	4	5	6	7	8
1	Simulationsuhr	0	12	14	15	22	24	27	28
2	kumulierte Laufzeit T_1	25	25	25	25		25		25
3	kumulierte Laufzeit T_2	12	12	29	29		29		29
4	kumuliertes Vorbeugungsintervall T_1	24	24	24	24		24		24
5	kumuliertes Vorbeugungsintervall T_2	15	15	15	33		33		30
6	bisherige Laufzeit A_1	0	12	12	13		22		
7	Produktionszeit bis Ausfall A_1	12-0		25-12	25-13				
8	Produktionszeit bis Vorbeugungszeitpunkt A_1	15-0		15-14	24-15				
9	kumulierte Laufzeit T_3	22				22		22	50
10	kumulierte Laufzeit T_4	33				33		33	33
11	kumuliertes Vorbeugungsintervall T_3	30				30		30	30
12	kumuliertes Vorbeugungsintervall T_4	27				27		27	55
13	bisherige Laufzeit A_2	0				22		22	22
14	Produktionszeit bis Ausfall A_2	22-0							33-22
15	Produktionszeit bis Vorbeugungszeitpunkt A_2	27-0							30-22
16	Anzahl eingesetzter Kräfte K_1	0	0	0	0	2	2	2	2
17	Anzahl eingesetzter Kräfte K_2	0	1	0	0	0	0	3	0

T_i = Reparaturzeit mit Nr. i ($i = 1, 2, 3, 4$)
 A_j = Produktionsanlage mit Nr. j ($j = 1, 2$)
 K_k = Instandhaltungskapazität mit Nr. k ($k = 1, 2$)

Im Zeitpunkt 15 wird zunächst geprüft, ob die bisherige Laufzeit des Teiles die Mindestlaufzeit erreicht hat. Die Mindestlaufzeit ist gleich dem Vorbeugungsintervall abzüglich der Dauer der vorbeugenden Reparatur multipliziert mit α , also $(18 - 2) \times 0,5 = 8$. Die bisherige Laufzeit von T_2 beträgt dagegen erst eine ZE, da dieses Teil ja im Zeitpunkt 14 gerade erst eingebaut worden ist. Die vorbeugende Reparatur von T_2 im Zeitpunkt 15 wird daher unterlassen. Der nächste Vorbeugungszeitpunkt wird durch das kumulierte Vorbeugungsintervall von T_2 (Zeile 5, Spalte 4) bestimmt, das gleich dem bisherigen kumulierten Vorbeugungsintervall (15) zuzüglich dem Vorbeugungsintervall (18) ist. Somit wird frühestens im Zeitpunkt 33 T_2 vorbeugend repariert (Zeile 5, Spalte 4). Aus dem Vergleich der kleinsten Restlaufzeit (Zeile 7, Spalte 4) und dem kleinsten Restvorbeugungsintervall (Zeile 8, Spalte 4) ergibt sich als nächstes Ergebnis für die Anlage 1 das Erreichen des Vorbeugungszeitpunktes von T_1 im Simulationszeitpunkt 24.

Im Zeitpunkt 22 fällt T_3 aus und wird von zwei Instandhaltungskräften (Zeile 16, Spalte 5) der ersten Instandhaltungsabteilung sechs Zeiteinheiten lang repariert.

Im Zeitpunkt 24 muß T_1 vorbeugend repariert werden, da die tatsächliche Laufzeit (22) die Mindestlaufzeit $(24 - 1) \times 0,5 = 11,5$ überschritten hat. Die Reparaturkapazität ist jedoch durch T_3 belegt, so daß T_1 bis zum Zeitpunkt 28 warten muß.

Im Zeitpunkt 27 wird mit der vorbeugenden Reparatur von T_4 begonnen. T_4 beansprucht drei Arbeitskräfte der Reparaturkapazität 2 (Zeile 17, Spalte 7).

Im Zeitpunkt 28 ist sowohl die Schadensreparatur von T_3 als auch die vorbeugende Reparatur von T_4 beendet. Die kumulierte Laufzeit von T_3 wird neu berechnet. Sie ist gleich der alten kumulierten Laufzeit (22) zuzüglich der neuen Laufzeit der zweiten Generation von T_3 (28) und beträgt somit gleich 50 (Zeile 9, Spalte 8). Das kumulierte Vorbeugungsintervall von T_4 beträgt jetzt $27 + 28 = 55$ (Zeile 12, Spalte 8). T_1 , das vom Zeitpunkt 24 an bis jetzt gewartet hat, kann nun vorbeugend

repariert werden. Es beansprucht zwei Arbeitskräfte der Instandhaltungsabteilung 1 (Zeile 16, Spalte 8). Die Anlage 2 nimmt die Produktion wieder auf und zwar bis zur vorbeugenden Reparatur von T_3 im Zeitpunkt 30.

Das Modell wurde exakt in der Simulationssprache GPSS/360 formuliert.⁴⁾ Es umfaßt etwa 150 Blöcke. Für den Fall, daß für alle Teile aller Anlagen die Feuerwehrstrategie verwendet wird, wurde, um Rechenzeit zu sparen, ein besonderes Modell entwickelt. Dafür genügten 40 Blöcke. Bei der Programmierung beanspruchte die Abbildung der Produktionsphase etwa 2/3 der Arbeitszeit. Schwierigkeiten macht vor allem die Abhängigkeit der Reparaturteile einer Anlage voneinander. Die Programmierung der Instandhaltungsphase war dagegen relativ einfach. Der Grund dafür ist, daß der Simulationssprache GPSS/360 ein allgemeines Warteschlangenmodell zugrundeliegt. Sie ist somit für die Abbildung der Warteschlangenproblematik in der Instandhaltungsphase prädestiniert, dagegen jedoch ein relativ schwerfälliges Instrument zur Abbildung der erneuerungstheoretischen Problematik in der Produktionsphase.

4 Modellexperimente

4.1 Experimental Design

Der Anlagenpark besteht aus 8 gleichartigen Anlagen mit jeweils 16 kritischen Teilen. Die Laufzeiten von 12 der Teile einer Anlage sind weibullverteilt mit steigender Ausfallrate⁵⁾ (hazard rate), die Laufzeiten der übrigen sind negativ exponential verteilt. Da die Ausfallrate der negativen Exponentialverteilung konstant ist, ist für diese Teile im Hinblick auf die oben genannten Zielsetzungen (s. Kapitel 2.5) die Feuerwehrstrategie optimal. Für die Teile mit weibullverteilten Laufzeiten werden dagegen neben der Feuerwehrstrategie 15 Vorbeugungsstrategien untersucht.

Die Instandhaltungsorganisation besteht aus zwei Abteilungen. Die erste kleinere Abteilung hält die Teile mit negativ exponentialverteilten und die

größere Abteilung die Teile mit weibullverteilten Laufzeiten instand. Die Reparaturzeiten für Schadensreparaturen der Teile sind logarithmisch normalverteilt. ⁶⁾

Es wird für dieses Produktions- und Instandhaltungssystem untersucht, wie Änderungen der Reparaturstrategien, der Reparaturkapazitäten und der Reparaturzeitverhältnisse die genannten Zielsetzungen beeinflussen. Die Einflußgrößen können folgende Ausprägungen aufweisen:

t, α -Strategien:

$\alpha = 0, 0,5, 1$ und $t = 40, 70, 100, 130, 160$ % der mittleren Laufzeit des Teiles. Jede t, α -Kombination stellt eine bestimmte Vorbeugungsstrategie dar. Die Vorbeugungsstrategien für die vorbeugend instandzuhaltenden Teile waren bei jedem Simulationslauf untereinander gleich. Für Teile mit negativ exponentialverteilten Laufzeiten wurde immer die Feuerwehrstrategie genommen. Die Anfangsintervalle wurden bei jeder Vorbeugungsstrategie so gewählt, daß einerseits die Vorbeugungszeitpunkte mehrerer Teile einer Anlage zusammenfielen, daß aber andererseits die für ihre Reparatur notwendige Instandhaltungskapazität die Maximalkapazität nicht überstieg.

Reparaturkapazitäten: (2,8), (3,12), (4,16). Die erste Zahl jedes Wertepaares entspricht der Zahl der Handwerker der ersten, die zweite der Zahl der Handwerker der zweiten Abteilung.

Reparaturzeitverhältnis: Verhältnis der Dauer der vorbeugenden Reparatur eines Anlagenteiles zur mittleren Dauer der Schadensreparatur des Teiles. 0,8 : 1; 0,6 : 1; 0,4 : 1.

Um den Einfluß der Größe des Produktions- und Instandhaltungssystems auf die untersuchten Zusammenhänge abschätzen zu können, wurde in einem ergänzenden Experiment seine Dimension verändert. Die Zahl der Reparaturteile der Anlagen wurde bei sonst gleichen Eigenschaften verdoppelt, ebenso die Anzahl der Arbeitskräfte der beiden Reparaturabteilungen. Das Reparaturzeitverhältnis betrug bei allen Läufen 0,4 : 1, insgesamt wurden somit bei diesem ergänzenden Experiment zusätzlich $270 : 3 = 90$ zusätzlich von 3×2 Simulationsläufen bei Anwendung der Feuerwehrstrategie vorgenommen.

4.2 Simulationsergebnisse

Stillstandsratenverläufe

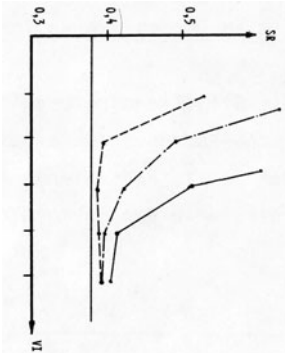
Die Abb. 1 zeigt Stillstandsratenverläufe. Auf der Ordinate jeder Abbildung ist die Stillstandsrate, auf der Abszisse das Vorbeugungsintervall abgetragen. Die Beobachtungswerte (jeweils der Mittelwert aus den beiden Simulationsläufen) sind als fette Punkte in dieses Achsenkreuz eingezeichnet. Sie sind durch gerade Linien verbunden. Die durchgezogene Linie gilt in allen Abbildungen für $\alpha = 0$, die punktiert gestrichelte für $\alpha = 1/2$ und die gestrichelte für $\alpha = 1$. Die Waagerechte zur Abszisse gibt die Stillstandsrate an, wenn für alle Teile der Anlagen die Feuerwehrstrategie genommen wird. Für nebeneinanderliegende Abbildungen ist die Instandhaltungskapazität, für untereinanderliegende Abbildungen das Reparaturzeitverhältnis gleich.

Aufgrund der Simulationsergebnisse sind folgende Zusammenhänge zu vermuten:

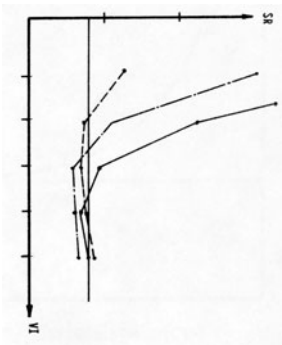
Vergleich: Feuerwehrstrategie - Vorbeugungsstrategien

1. Bei einer ausreichenden Verringerung der Reparaturzeitverhältnisse werden die Stillstandsrate für Vorbeugungsstrategien unter das Niveau für die Feuerwehrstrategie gesenkt. Die Größe des Zuwachses im Unterschied zwischen den Stillstandsrate beider Strategieklassen hängt wesentlich von dem Ausmaß der Herabsetzung des Reparaturzeitverhältnisses ab, nicht dagegen von dem Reparaturzeitverhältnis vor der Herabsetzung oder von der Größe der Instandhaltungskapazität.
2. Bei einer Erweiterung der Instandhaltungskapazität werden die Stillstandsrate der Vorbeugungsstrategien grundsätzlich stärker gesenkt als die der Feuerwehrstrategie. Bei prozentual gleicher Kapazitätsausdehnung ist der Zuwachs im Unterschied zwischen den Stillstandsrate der beiden Strategieklassen jedoch umso geringer, je größer die Kapazität und je kleiner das Reparaturzeitverhältnis vor der Kapazitätsvergrößerung war.
3. Die Aussagen lassen sich umkehren für eine Verringerung der Reparaturkapazität und bei einer Vergrößerung des Reparaturzeitverhältnisses.

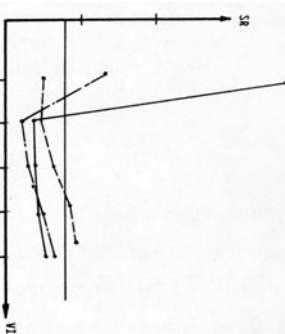
Reparaturzeitverhältnis 0,8 : 1



Reparaturzeitverhältnis 0,6 : 1

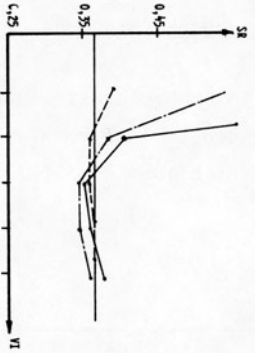


Reparaturzeitverhältnis 0,4 : 1



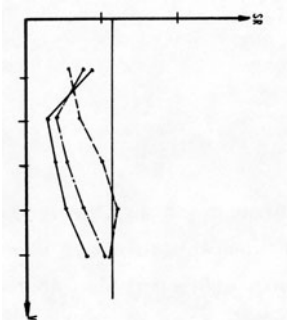
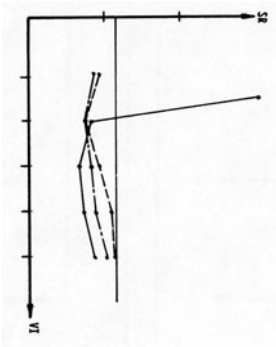
Instandhaltungskapazität 8t

(2,8)



Instandhaltungskapazität 18t

(3,12)



Instandhaltungskapazität 18t

(4,16)

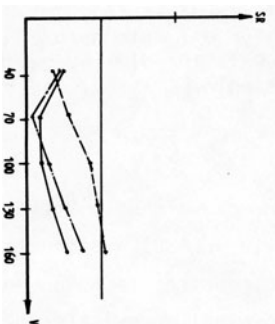
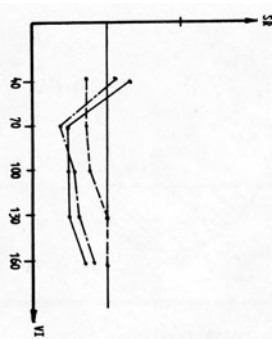
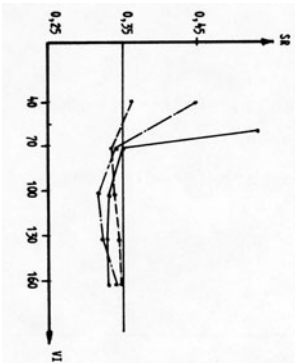


Abb. 1: Stillstandsratemerkmale

Vergleich verschiedener Vorbeugungsstrategien

1. Die Stillstandsrate weist in Abhängigkeit vom Vorbeugungsintervall zwei typische Kurvenverläufe auf – den Verlauf vom Typ 0 und den vom Typ U. Bei beiden Stillstandsratenverläufen bildet die Stillstandsrate für die Feuerwehrstrategie die Asymptote.

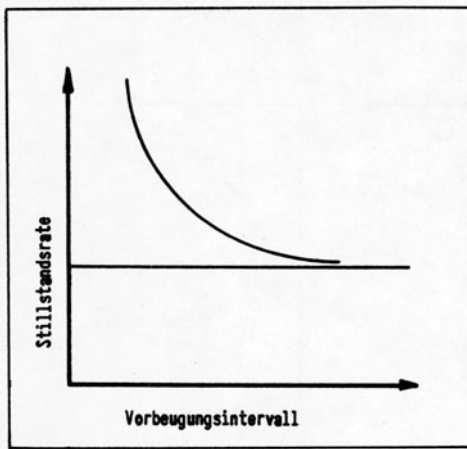


Abb. 2: Stillstandsratenverlauf vom Typ 0 (oberhalb der Rate für die Feuerwehrstrategie)

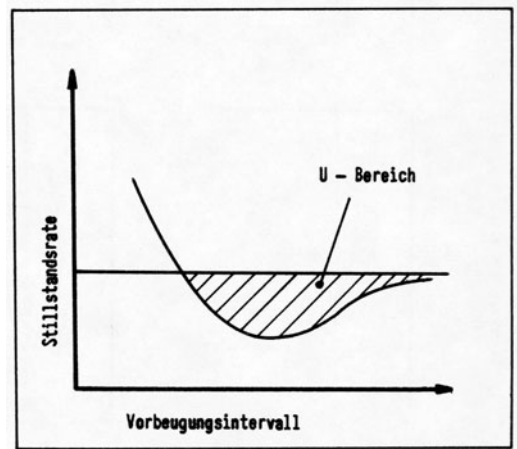


Abb. 3: Stillstandsratenverlauf vom Typ U (unterhalb der Rate für die Feuerwehrstrategie)

2. Aufgrund der Simulationsergebnisse ist zu vermuten, daß sich die Stillstandsratenverläufe vom Typ U in Abhängigkeit von α wie in Abb. 4 dargestellt verändern. Der Kurvenzug links außen hat den höchsten, der rechts außen den niedrigsten α -Wert. Danach würde eine optimale t, α -Kombination existieren.

3. Eine Verkleinerung des Reparaturzeitverhältnisses und/oder eine Kapazitätserweiterung vergrößern den U-Bereich des Stillstandsratenverlaufs für ein bestimmtes α . Ferner setzen sie das optimale Vorbeugungsintervall herab.

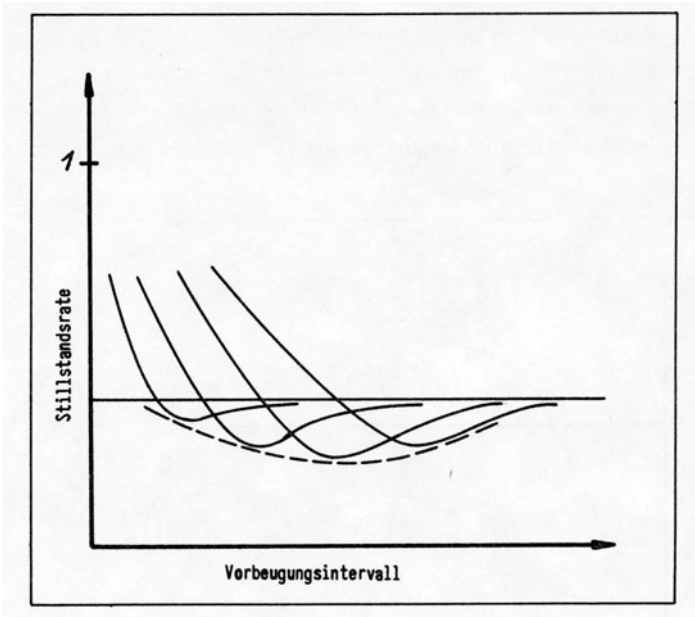


Abb. 4 Stillstandsratenverlauf in Abhängigkeit von Parameter α

4. Die genannten Auswirkungen einer Kapazitätserweiterung sind jedoch umso geringer, je kleiner das Reparaturzeitverhältnis und je größer die Kapazität vor der Kapazitätsvergrößerung waren.

Auslastungsgrade der Reparaturkapazitäten

Neben den Stillstandsraten wurden bei den oben genannten Experimenten auch die Auslastungsgrade der beiden Reparaturkapazitäten gemessen – der Reparaturkapazität für Teile mit negativ exponentialverteilten Laufzeiten (Nicht-Verschleißteile) und der Reparaturkapazität für Teile mit weibullverteilten Laufzeiten (Verschleißteile). Die Simulationsergebnisse legen den in Abb. 5 dargestellten Verlauf für den Auslastungsgrad der Reparaturkapazität für Teile mit negativ exponentialverteilten Laufzeiten nahe. Bei kleinen Vorbeugungsintervallen für die Verschleißteile ist der Auslastungsgrad der Reparaturkapazität für die Nicht-Verschleißteile gering.

Er steigt mit wachsendem Vorbeugungsintervall und erreicht ein Maximum oberhalb des Auslastungsgrades, der sich einstellt, wenn für alle Teile die Feuerwehrstrategie angewendet wird. Danach fällt er ab und nähert sich diesem Niveau um so mehr, je größer die Vorbeugungsintervalle werden.

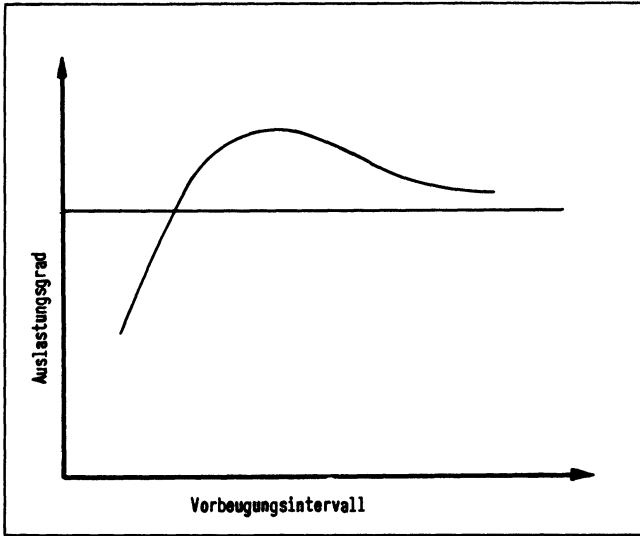


Abb. 5 Typischer Verlauf des Auslastungsgrades der Reparaturkapazität für Nicht-Verschleißteile

Bei einem Vergleich mit dem Verlauf der Stillstandsrate fällt auf, daß sie sich weitgehend entgegengesetzt zueinander entwickeln. Wenn die Stillstandsrate hoch ist, ist der Auslastungsgrad der Reparaturkapazität für Nicht-Verschleißteile niedrig und umgekehrt. Dieses entgegengesetzte Verhalten läßt sich plausibel erklären. Ist die Stillstandsrate hoch, so ist in einem bestimmten Planungszeitraum der Verbrauch an Nicht-Verschleißteilen, bei denen die Laufzeit ja immer voll ausgenutzt wird, gering und damit auch der Auslastungsgrad der Reparaturkapazität für diese Teile. Liegt dagegen die Stillstandsrate unter derjenigen für die Feuerwehrstrategie, so verstreichen auch die Laufzeiten der Nicht-Verschleißteile in kürzerer Kalenderzeit als bei der Feuerwehrstrategie oder auf einen festen Planungszeitraum bezogen: Es werden mehr Nichtverschleißteile verbraucht als bei der Feuerwehrstrategie. Der Auslastungsgrad der Reparaturkapazität für diese Teile liegt dementsprechend höher als bei Anwendung der Feuerwehrstrategie (für alle Teile).

Minimierung der Stillstandsrate durch Wahl geeigneter Vorbeugungsstrategien für Verschleißteile bedeutet somit vermutlich gleichzeitig Maximierung der Beanspruchung der Reparaturkapazität für Nichtverschleißteile.

Der Auslastungsgrad der Reparaturkapazität für Verschleißteile verläuft in Abhängigkeit vom Vorbeugungsintervall im Prinzip ähnlich wie die Stillstandsrate. Jedoch liegt die minimale Stillstandsrate in allen Fällen bei einem kleineren Vorbeugungsintervall als der minimale Auslastungsgrad der Reparaturkapazität für Verschleißteile. Zwischen beiden Minima steigt die Stillstandsrate und nähert sich dem Wert für die Feuerwehrstrategie, während die Kapazitätsausnutzung sinkt. Diese entgegengesetzte Entwicklung kann man wie folgt erklären:

Beim optimalen Vorbeugungsintervall bezogen auf die Stillstandsrate ist die Summe aus den Zeiten für vorbeugende Reparaturen und Zufallsreparaturen sowie aus den Wartezeiten aller Anlagen minimal. Mit wachsendem Vorbeugungsintervall sinkt, da das Minimum der Kapazitätsausnutzung für Verschleißteile noch nicht erreicht ist, die gesamte Reparaturzeit für Verschleißteile, da der maximale Auslastungsgrad der Reparaturkapazität für diese Teile überschritten wurde. Der Anstieg der Stillstandsrate resultiert dann einzig aus der zunehmenden Unregelmäßigkeit in der Reparaturterminierung aufgrund des wachsenden Anteils der Zufallsreparaturen an der Gesamtzahl der Reparaturen.

Der Entscheidende muß bei der Festlegung des Vorbeugungsintervalls in diesem Bereich zwischen geringerer Kapazitätsauslastung und damit eventuell geringeren Instandhaltungskosten und höherer Instandhaltungsrate und damit eventuell geringeren Erlösen aus der Produktion abwägen.

Ersatzteilverbrauch

Der Ersatzteilverbrauch ist bei der Feuerwehrstrategie (für alle Teile) immer am niedrigsten. Bei Anwendung von Vorbeugungsstrategien fällt er mit

wachsendem Vorbeugungsintervall und nähert sich dem Verbrauch für die Feuerwehrstrategie. Für alle Wertkonstellationen der übrigen Faktoren liegt der Ersatzteilverbrauch für $\alpha = 1$ unter jenem für $\alpha = 1/2$ und diese wiederum unter dem für $\alpha = 0$. Unterteilt man den Gesamtverbrauch in Verbrauchsmengen für vorbeugende Reparaturen und für Schadensreparaturen so stellt man fest, daß mit wachsendem α ceteris paribus die Anzahl der Schadensreparaturen zwar zunimmt, sie wird jedoch überkompensiert durch die Abnahme der vorbeugenden Reparaturen, so daß der Gesamtverbrauch fällt. Die Senkung des Reparaturzeitverhältnisses und die Kapazitätserweiterung wirken auf den Ersatzteilverbrauch gerade umgekehrt wie auf die Stillstandsrate, sie erhöhen ihn, und zwar umso stärker, je kleiner das Vorbeugungsintervall ist.

Die Dimension des Produktionsbetriebes und der Instandhaltungsorganisation (Anzahl der Reparaturteile; Anzahl der Instandhaltungskräfte) wurden bei den vorliegenden Simulationsstudien gegenüber der Realität verkleinert. In einer weiteren Gruppe von Experimenten wurde die Anzahl der Reparaturteile und die Anzahl der Instandhaltungskräfte erhöht. Dabei stellte sich heraus, daß die obigen Aussagen aufrechterhalten werden konnten.

5 Modellerweiterungen

Erweiterungen um ein Produktionsmodell: Die einzelnen Fertigungsanlagen des Produktionsbetriebes (nicht die Teile einer Anlage) wurden bisher als voneinander unabhängig behandelt. Sind die Anlagen jedoch tatsächlich durch einen Produktionsfluß miteinander verknüpft, so werden im Rahmen des vorliegenden Modells Aussagen über instandhaltungsbedingte Produktionsausfälle erschwert. Eine Möglichkeit wäre z. B. die Produktionsausfälle in Abhängigkeit von den Ausfallzeiten der Engpaßeinheit festzustellen. Dies ließe sich mit relativ geringfügigen Modelländerungen realisieren. Man könnte jedoch auch versuchen, den Fluß der Produktionsaufträge durch das System von Fertigungsanlagen im Modell selbst abzubilden. Dies würde jedoch umfangreiche Modelländerungen erfordern.

Instandhaltungsorganisation: Da das Simulationsmodell in der Sprache GPSS/360 formuliert wurde und da diese für Abbildung von Warteschlangenproblemen besonders prädestiniert ist, können mit relativ geringfügigen Modelländerungen anders strukturierte Instandhaltungsorganisationen abge-

bildet werden; z. B. könnte man die Regelung, daß für Schadensreparaturen und für vorbeugende Reparaturen eines bestimmten Reparaturteils immer nur eine Instandhaltungsabteilung zuständig ist, abändern. Man könnte getrennte Instandhaltungsabteilungen für vorbeugende und Schadensreparaturen einführen; man könnte einfügen, daß ein Reparaturteil mehrere Instandhaltungsabteilungen alternativ, gleichzeitig oder sukzessiv beanspruchen kann.

Prioritätsregeln: Die Simulationssprache GPSS/360 ist im Hinblick auf Variationen der Prioritätsregelung besonders flexibel konstruiert. Mit geringfügigen Modelländerungen können auch beliebige andere Prioritätsregeln abgebildet werden.

Leerzeitstrategien: Bisher wurde von kontinuierlicher Produktion ausgegangen, die lediglich durch Instandhaltungsbedingte Warte- und Reparaturzeiten unterbrochen wurde. Bei Auftreten von produktionsbedingten Leerzeiten wird man die Vorbeugungszeitpunkte in diese Zeiten legen, um die Produktionsausfälle zu senken. Regelmäßig auftretende Leerzeiten (Wochenenden, dritte Schicht) können im Modell leicht berücksichtigt werden. Man braucht lediglich die Vorbeugungsintervalle der t, α -Strategien so zu wählen, daß die Vorbeugungszeitpunkte in die Leerzeiten fallen und die Warte- und Reparaturzeiten innerhalb der Leerzeiten bei der Berechnung der instandhaltungsbedingten Stillstandszeiten wegzulassen.

Instandhaltungsziele: Ähnlich wie in der Realität ist es auch im vorliegenden Simulationsmodell im Prinzip möglich, jeden einzelnen Schritt der Reparaturteile während der Produktions- und Instandhaltungsphase aufzuzeichnen und daraus über die hier festgehaltenen drei Zielgrößen hinaus weitere Zielgrößen zu ermitteln (z. B. durchschnittliche Wartezeiten der Instandhaltungsaufträge, durchschnittliche Warteschlangenlänge, mittlere Länge der Produktionszeiten usw.). Ferner kann man das Zielkonzept dahingehend erweitern, daß man einige der wesentlichen hier gemessenen Zielgrößen in den Kostenraum oder den Gewinnraum abbildet, um somit zu eindeutigen Aussagen über die Vorteilhaftigkeit bestimmter Reparaturstrategien zu gelangen. Man könnte z. B. die Stillstandszeiten mit den entgangenen Gewinnen gewichten. Ferner den Ersatzteilverbrauch mit den Ersatzteilpreisen und eventuell einem Lagerkostenzuschlag und die

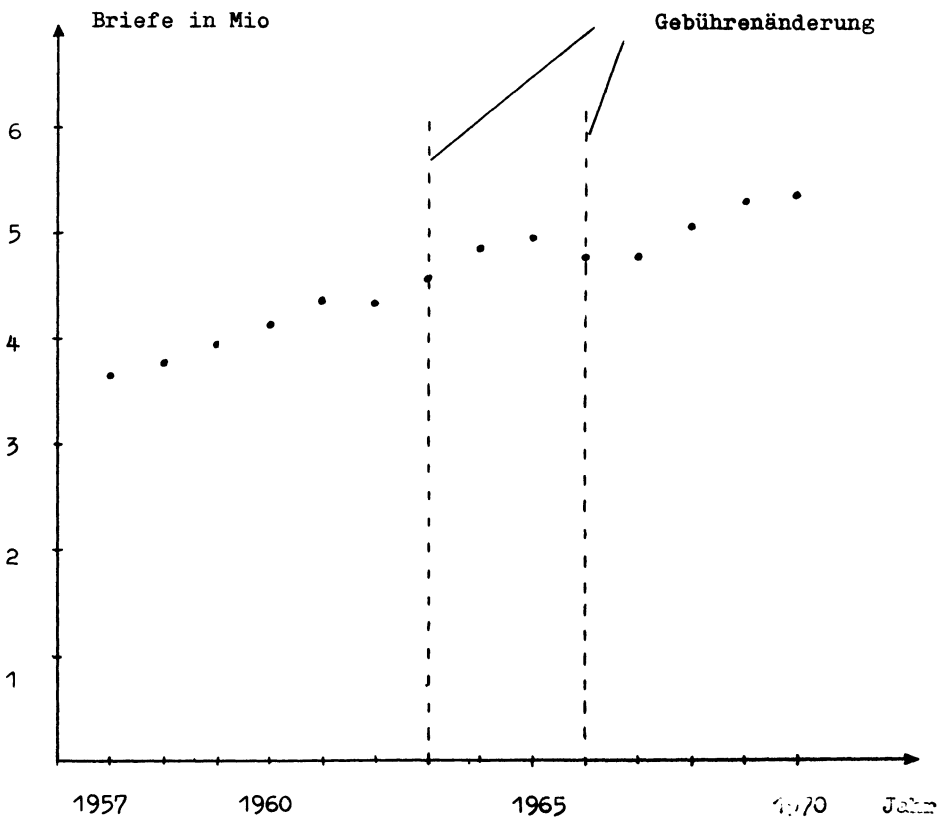
Arbeitszeiten der eingesetzten Handwerker mit bestimmten Lohnsätzen. Die bisher nicht komparablen Zielgrößen wären somit komparabel und könnten zu einer Größe Instandhaltungskosten (einschließlich der Stillstandsverluste) zusammengefaßt werden.

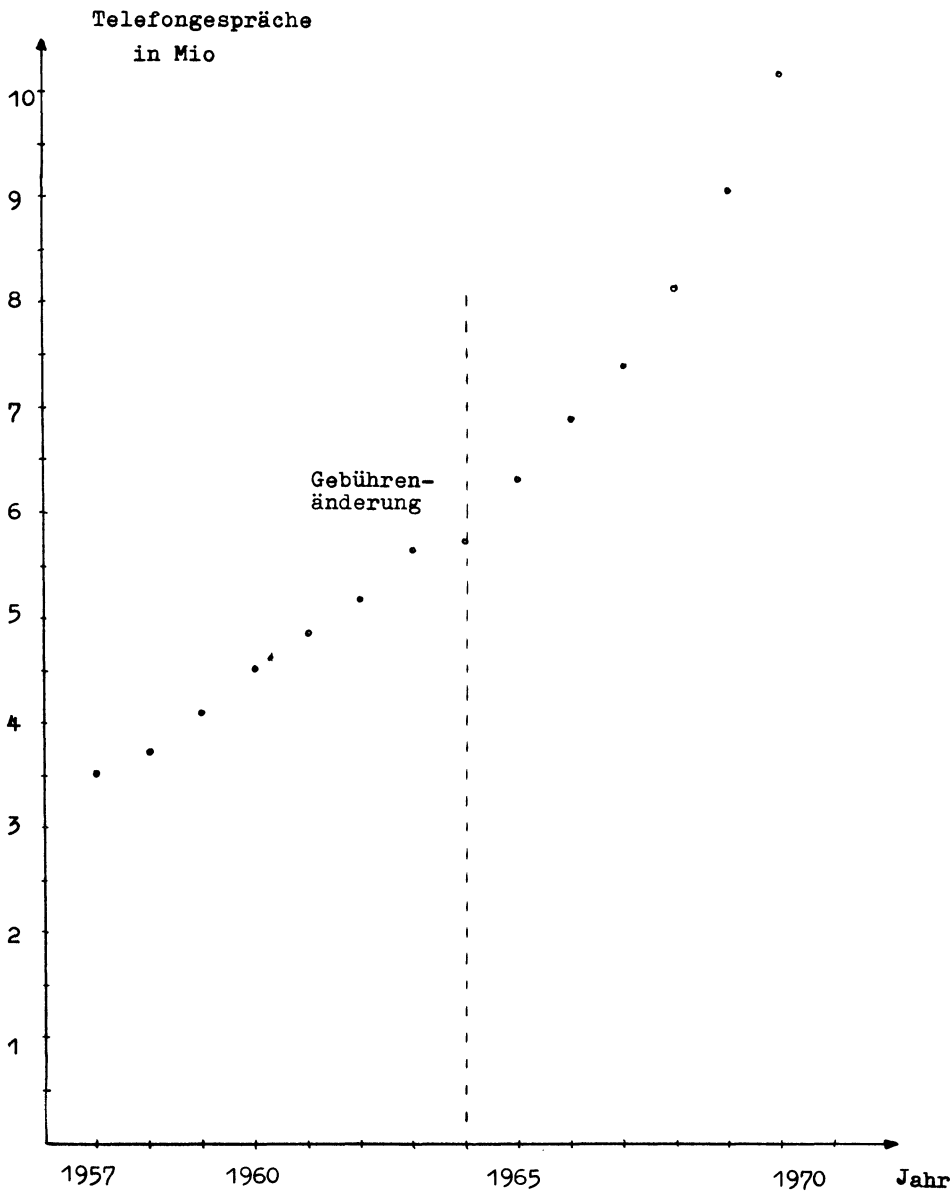
- 1) Vergin, R. C.: Scheduling Maintenance and Determining Crew Size for Stochastically Failing Equipment, in: Management Science, vol. 13, 1966, S. B-52 - B-65.
- 2) Frotscher, J.: Simulationsmodell für den Reparaturdienst mit Berücksichtigung von Wegezeiten, in: Rechentechnik/Datenverarbeitung, Bd. 4, 1967, Heft 12, S. 14 - 19; derselbe: Ein Simulationsmodell für den Reparaturdienst, in: Simulationsmodelle für ökonomische - organisatorische Probleme, hrsg. vom Institut für Datenverarbeitung -Dresden, Köln und Opladen 1968, S. 127 - 169.
- 3) Kress, H.: Untersuchungen zur Bestimmung der optimalen Organisation von Instandhaltungsarbeiten an Fertigungsmaschinen bei Werkstättenfertigung anhand eines Simulationsmodells, Diss. München 1968; Kurzfassung in: Unternehmensforschung, Bd. 12, 1968, S. 269 - 280.
- 4) Vgl. im einzelnen Ordelheide, Dieter: Instandhaltungsplanung, Simulationsmodelle für den Instandhaltungsbereich, Wiesbaden 1973 (in Vorbereitung).
- 5) Es sei $F(t)$ die Wahrscheinlichkeitsverteilung der Laufzeit eines Reparaturteils und $f(t)$ die Dichte der Wahrscheinlichkeitsverteilung, dann ist $f(t) / (1-F(t))$ die Ausfallrate. Man kann sie intuitiv erklären als Wahrscheinlichkeit eines im Zeitpunkt t unmittelbar bevorstehenden Eintritts der Funktionsunfähigkeit eines Reparaturteils unter der Annahme, daß es das Alter t bereits erreicht hat. Vgl. im einzelnen z. B. Cox, D. R.: Erneuerungstheorie, München 1965, S. 13 ff.
- 6) Zur Wahl der Weibullverteilung und der logarithmischen Normalverteilung vgl. Kress, H.: a.a.O., S. 13 ff.; Ordelheide, Dieter: a.a.O., und die dort zitierte Literatur.

Makroökonomische Anwendungen

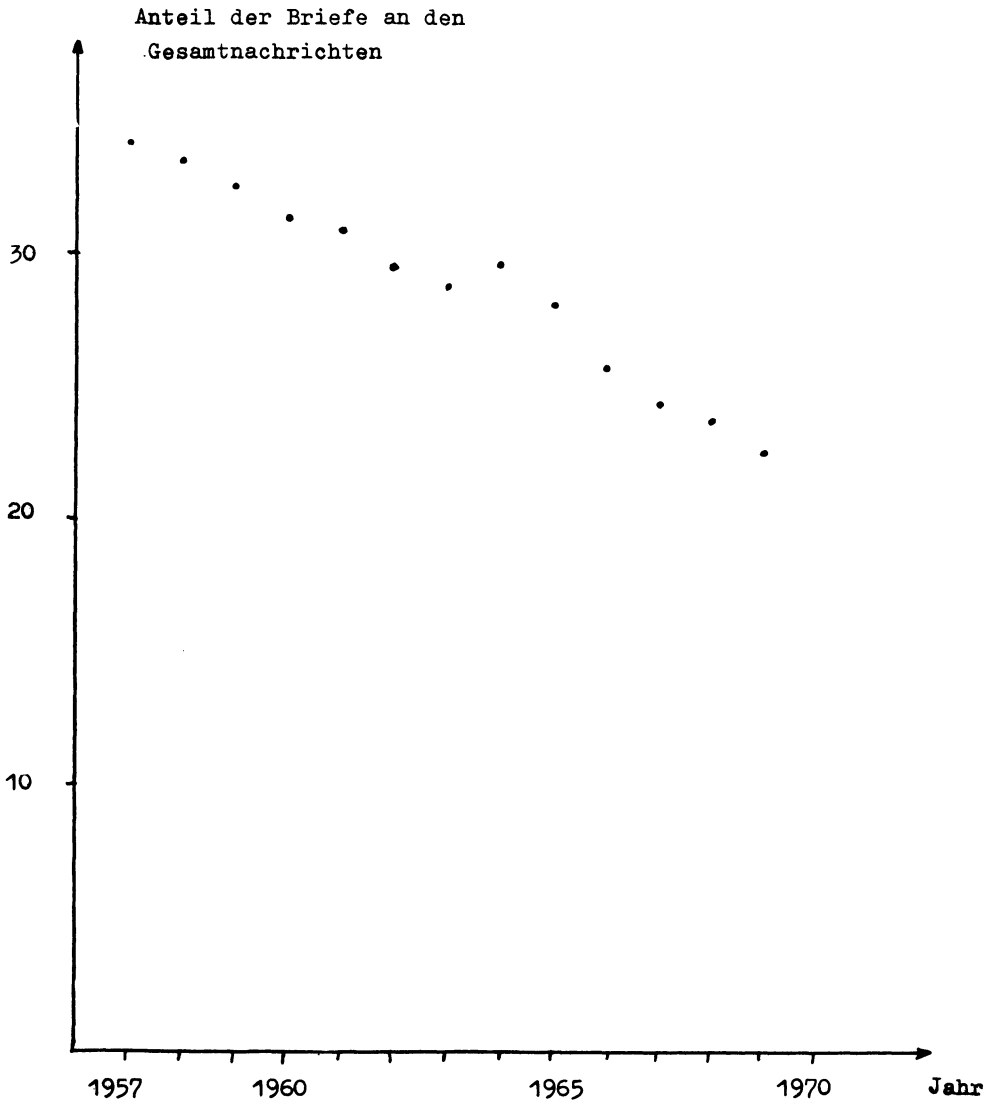
**Zwischenbericht über eine Analyse des Zusammenhanges zwischen
Brief- und Fernsprechverkehr im Bereich der Deutschen Bundespost**
von U. Becker, Darmstadt

Es soll hier über die vorbereitenden Arbeiten für eine Prognose des zu erwartenden Nachrichtenverkehrs bei der DEUTSCHEN BUNDES-POST berichtet werden. Der erste Schritt jeder Vorausschätzung besteht darin, die Entwicklung in der Vergangenheit zu betrachten. Wir haben Unterlagen über die per Post und die per Fernsprecher ausgetauschten Nachrichten. In den letzten vierzehn Jahren sah die Entwicklung folgendermaßen aus:





Sowohl bei den Gesprächen, als auch bei den Briefen ist eine Aufwärtsentwicklung zu beobachten. Dabei ist die Wachstumsrate für Gespräche größer als für Briefe. Demnach ist die langfristige Entwicklung des Anteils der Briefe an den gesamten Nachrichten in dem dargestellten Zeitraum rückläufig.



Für unsere weiteren Überlegungen halten wir fest, daß die beiden Kommunikationsmittel Brief und Telefongespräch nicht in vollem Maße austauschbar sind. Es ist kaum vorstellbar, daß in absehbarer Zeit die Übersendung von Rechnungen, Schecks usw. durch Telefongespräche ersetzt werden kann. Umgekehrt kann ein eiliger Telefonanruf nicht durch einen Brief ersetzt werden. Zwischen diesen Grenzfällen reicht die Skala von einem Telefongespräch zur Übermittlung einer einseitigen Nachricht, auf welche keine Antwort benötigt wird (d.h. ein Telefongespräch entspricht einem Brief), über ein Telefongespräch, welches eine unmittelbare Antwort erforderlich macht (d.h. ein Telefongespräch entspricht zwei Briefen), bis zu einer langen Unterhaltung, die aus aufeinanderfolgenden Fragen und Antworten besteht (d. h. ein Telefongespräch entspricht mehreren Briefen). Über die Häufigkeit dieser verschiedenen Äquivalente existieren keine Unterlagen. Für den Zweck dieser Arbeit wurde davon ausgegangen, daß durchschnittlich ein Telefongespräch zwei Briefen entspricht. Damit folgt, daß die Gesamtnachrichten sich zusammensetzen aus der Zahl der Briefe zuzüglich der zweifachen Zahl der Telefongespräche.

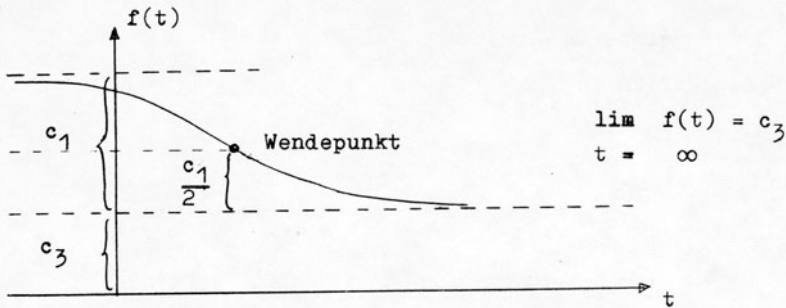
Begnügen wir uns vorerst mit einer Trendextrapolation, dann erhalten wir die Prognosewerte durch autoregressive Verlängerung einer theoretischen Funktion, die dem gegebenen Datenmaterial angepaßt wurde. Für die Zeitreihe des Anteils der Nachrichten, die durch den Postdienst übermittelt werden, wählen wir aus sachlogischen Überlegungen eine logistische Kurve von der Form:

$$A_B = \frac{C_1}{1 + C_2 \cdot \exp(\lambda \cdot t)} + C_3$$

A_B = Anteil der Briefe an den Gesamtnachrichten

t = Zeit

Die obere und die untere Grenze der Kurve werden durch die Parameter C_1 und C_3 bestimmt.



Wenn sich c_2 erhöht, verschiebt sich die Kurve nach rechts und wenn c_1 sich verringert, wird die Kurve flacher.

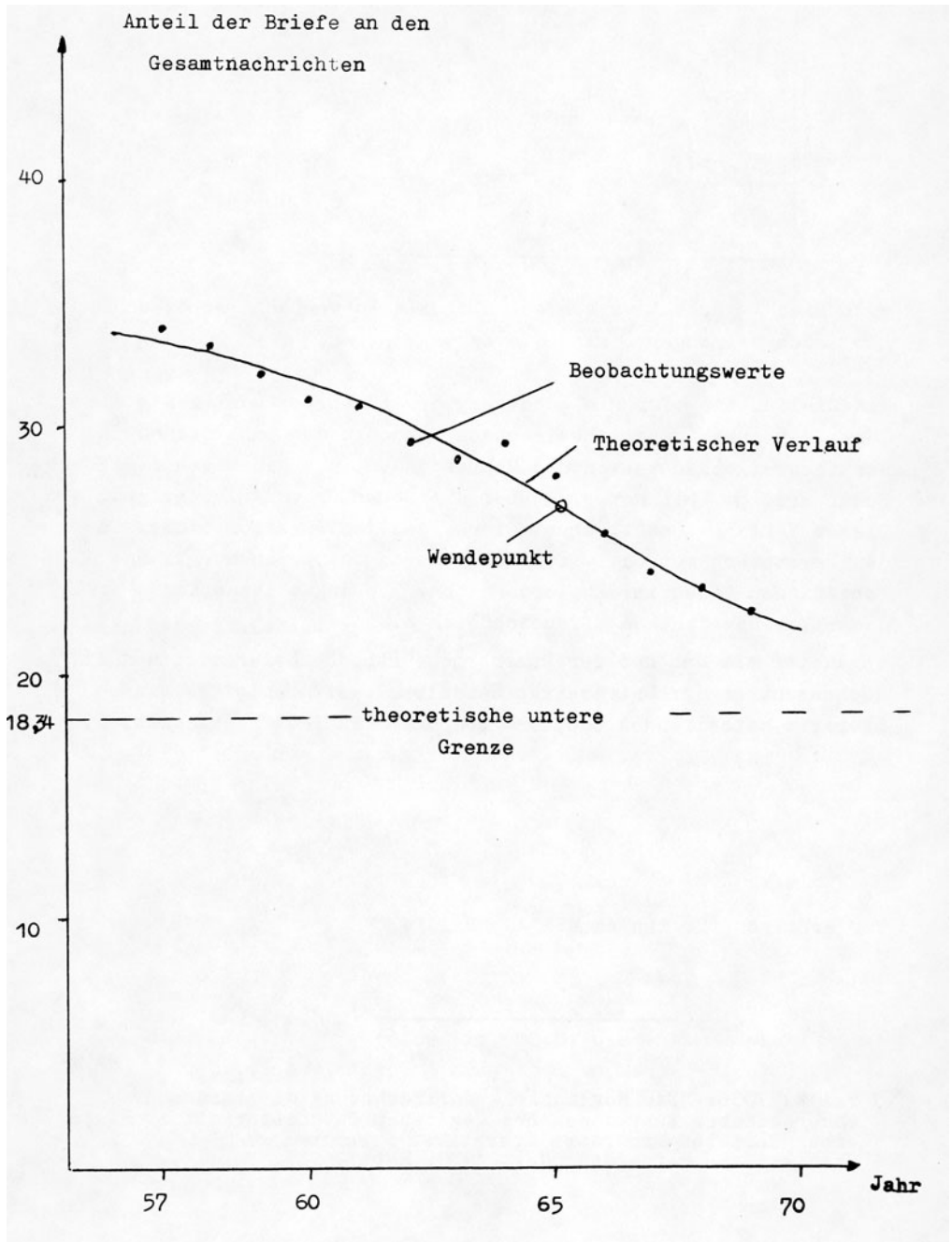
Als Zielfunktion für die Schätzung der Parameter wählen wir die Summe der quadrierten Abweichungen zwischen den empirischen und den theoretischen Werten des Regressanden. Die Minimierung dieser Summe erfolgt üblicherweise über die sogen. Normalgleichungen. Dieses Verfahren setzt aber voraus, daß die Funktion linear in den Parametern ist oder zumindest durch entsprechende Transformationen linearisiert werden kann. Die logistische Kurve erfüllt diese Voraussetzungen offensichtlich nicht. Deshalb wurden die Parameter mit dem auf der Rosenbrock-Methode basierenden und im Rechenzentrum der Universität Heidelberg weiterentwickelten iterativ arbeitenden Coputer-Programm "PAMIRO" berechnet. ⁺)

Wir erhalten für den Anteil der Briefe

$$A_B = \frac{17,12}{1 + \exp(0,26 \cdot t - 9,15)} + 18,34 \quad (\%)$$

$$t = 1, 2, \dots$$

⁺) Vgl W. VOSS: "Die Möglichkeit der Schätzung der Parameter theoretischer Funktionen bei gegebenem Datenmaterial mit Hilfe eines Computer-Programms iterativer Aproximation", Statistische Hefte, 11. Gg., 1970, Heft 2



Die zu prognostizierende Variable wird als Funktion der Zeit aufgefaßt. Damit hat man einen Regressor zur Hand, der mit Sicherheit in die Zukunft zu projizieren ist.

Der Regressand "Anteil der Briefe an den gesamten Nachrichten" hat eine theoretische untere Grenze von

18,34 %.

Eine Möglichkeit, zu einer Prognose zu kommen ist, die Funktion zu extrapolieren, unter der Annahme, daß die Tendenz der Einflußfaktoren sich nicht grundsätzlich ändert.

Die andere Möglichkeit: Man versucht die Einflußfaktoren qualitativ zu erkennen und die Abhängigkeit der zu untersuchenden Größe von diesen Einflußfaktoren quantitativ zu erfassen.

Für eine Prognose zu einem bestimmten Zeitpunkt setzt man dann nicht mehr direkt den Zeitpunkt ein, wie zuvor, sondern die für diesen Zeitpunkt prognostizierten Einflußfaktoren. Man hat mithin die Möglichkeit zusätzlich Informationen über die Entwicklung der Einflußfaktoren zu verwenden.

Entscheidender Einflußfaktor in unserer Untersuchung ist der Bestand an Hauptanschlüssen. Diese Größe hängt außer von der Bedarfsentwicklung vor allem von den Leistungen der Deutschen Bundespost ab, deren Ziel es sein muß, die Lücke zwischen Bestand und Bedarf möglichst bald zu schließen.

Über die zukünftige Bedarfsentwicklung der Zahl der Hauptanschlüsse haben bereits umfangreiche Untersuchungen von Seiten der Deutschen Bundespost stattgefunden, deren Ergebnisse uns vorliegen. Danach wird die Zahl der Hauptanschlüsse in der Bundesrepublik in 20 Jahren voraussichtlich auf das 3 bis 4fache anwachsen. Zu den 7,6 Mio Hauptanschlüssen, die bis Ende 1969 vorhanden waren, werden rund 20 Mio hinzukommen.

GERFIN meint zwar skeptisch:

"Bedenken sollte man aber auch, daß sich andere, raffiniertere Verfahren von den reinen Trendverfahren vielfach dadurch unterscheiden, daß die gesuchte Größe nicht unmittelbar extrapoliert wird, sondern in bestimmte Faktoren aufgebroschen wird, deren Entwicklung dann ihrerseits "naiv" extrapoliert wird."

Dieser Vorbehalt kann das hier zu erstellende Modell nicht treffen, zumindest nicht, was die Einflußgröße "Bestand an Hauptanschlüssen" angeht.

Das vorausgesagte starke Anwachsen der Zahl der Hauptanschlüsse ist der Hauptfaktor für die Entwicklung des Gesprächsaufkommens. Dieser Faktor ist auch zum großen Teil für die Verlagerung von Briefen auf Telefongespräche verantwortlich.

Neben dem Ausbau des Fernsprechsystems muß unser Modell weitere Einflußfaktoren, wie Gebührenänderungen und wirtschaftliches Wachstum berücksichtigen.

Es gibt Anhaltspunkte dafür, daß der Zuwachs der geschäftlichen Nachrichten sich von demjenigen des privaten Sektors unterscheidet. Es sind zwar Untersuchungen in dieser Richtung geplant, bisher gibt es jedoch keine gesonderten Unterlagen für geschäftliche und private Nachrichten, nur Gesamtzahlen für Briefe und Telefongespräche.

So erzwingen die praktischen Gegebenheiten schon beim Entwurf des Modells einen Kompromiß. Auf die Einbeziehung von Größen, die theoretisch noch so überzeugend wären, muß verzichtet werden, weil ihre Zahlenwerte statistisch nicht erfaßt sind. Andererseits können und sollen die bei der Konstruktion von Modellen gewonnenen Erkenntnisse Rückwirkungen auf die statistische Datenerfassung haben und dazu führen, daß künftig die Erfordernisse von Prognosemodellen bei der Erhebung berücksichtigt werden.

Versuchen wir nun ein einfaches Modell für die Entwicklung des Zusammenhanges zwischen Brief- und Fernsprechverkehr zu entwerfen.

Es bietet sich an, die Zahl der Telefongespräche aufzuspalten, so daß wir eine Gleichung für die Ortsgespräche und eine für die Ferngespräche bekommen. Das scheint deshalb erforderlich, weil die Einflußfaktoren verschieden sind. Bei einer Prognose muß allerdings berücksichtigt werden, daß durch die geplante Einführung des Nahtarifs in etwa 5 Jahren eine Verschiebung zwischen den Ortsgesprächen und den Ferngesprächen eintreten kann.

Als weitere Verhaltensgleichung wählen wir den Anteil der Briefe an den gesamten Nachrichten.

Was den Einflußfaktor "Zahl der Hauptanschlüsse" angeht, so wird sich die Zunahme der Hauptanschlüsse im Wesentlichen auf dem Sektor der Privathaushalte abspielen, denn der Bedarf an Geschäftsanschlüssen kann im wesentlichen als abgedeckt betrachtet werden. Somit erscheint es vernünftiger, als Einflußgröße nicht unmittelbar die Zahl der Hauptanschlüsse zu wählen, sondern die Zahl der Hauptanschlüsse dividiert durch die Zahl der Haushalte. Mit wachsender Anschlußdichte nimmt die Zahl der Gespräche zu. Damit geht automatisch der Anteil an Briefen zurück und zwar nicht nur deshalb, weil der Nenner wächst, sondern auch, weil durch die größere Anschlußdichte ein Teil des Briefverkehrs durch Telefongespräche ersetzt wird.

Für die Telefongespräche kommt als weiterer Einflußfaktor die jeweilige Gesprächsgebühr hinzu. Theoretische Überlegungen, und praktische Erfahrungen führen zu der Erwartung, daß eine Anhebung der Gesprächsgebühren zu einem Rückgang der Zahl der Telefongespräche und auch zu einem Überwechseln von Telefongesprächen zu Briefen führen können. Da steigende Preise bei Konstanz der nominalen Gebühren eine Senkung der realen Gebühren bedeutet, wählen wir als Einflußfaktor die Größe: Gesprächsgebühr dividiert durch Einzelhandelspreisindex.

Für die Zahl der Ferngespräche und für den Anteil der Briefe an den gesamten Nachrichten wird die Briefgebühr dividiert

durch den Einzelhandelspreisindex einen Einfluß haben.

Als volkswirtschaftliche Einflußgrößen wählen wir bei den Telefongesprächen den Einzelhandelsumsatz und bei den Ferngesprächen sowie bei dem Anteil der Briefe das Volkseinkommen.

Damit erhalten wir für die Verhaltensgleichungen:

$$\left| \begin{array}{l} N_o = f_1 (D ; GG_o/J ; E) \\ N_f = f_2 (D ; GG_f/J ; GB/J ; E ; V) \\ A_B = f_3 (D ; GG/J ; GB/J ; V) \end{array} \right.$$

N_o = Zahl der Ortsgespräche

N_f = Zahl der Ferngespräche

A_B = Anteil der Briefe an den
Gesamtnachrichten

D = Zahl der Hauptanschlüsse dividiert durch die Zahl der Haushalte

J = Einzelhandelspreisindex

GG_o = Gebühr für ein Ortsgespräch

GG_f = Durchschnittsgebühr für ein Ferngespräch

GG = Durchschnittsgebühr für ein Telefongespräch

GB = Durchschnittsgebühr für einen Brief

E = Einzelhandelsumsatz

V = Volkseinkommen

Wir vervollständigen das Modell durch zwei Definitionsgleichungen:

$$\left| \begin{array}{l} N_G = N_B + 2 \cdot (N_o + N_f) \\ A_B = \frac{N_B}{N_G} \end{array} \right.$$

N_G = Zahl der Gesamtnachrichten

N_B = Zahl der Briefe

Im bisher betrachteten Modell werden in den Verhaltensgleichungen eindeutige funktionale Beziehungen zwischen den Variablen unterstellt. Das ist offenbar sehr unrealistisch. In Wahrheit werden die endogenen Variablen noch durch unübersehbar viele Faktoren beeinflusst, welche nicht im Modell repräsentiert sind. All diese modellexternen Faktoren werden summarisch berücksichtigt durch die latenten Variablen u , v und w . Im einfachsten Fall in der Weise, daß eine additive Überlegung postuliert wird.

$$N_o = f_1 + u$$

$$N_f = f_2 + v$$

$$A_B = f_3 + w$$

Die latenten Variablen sind nach MENGES

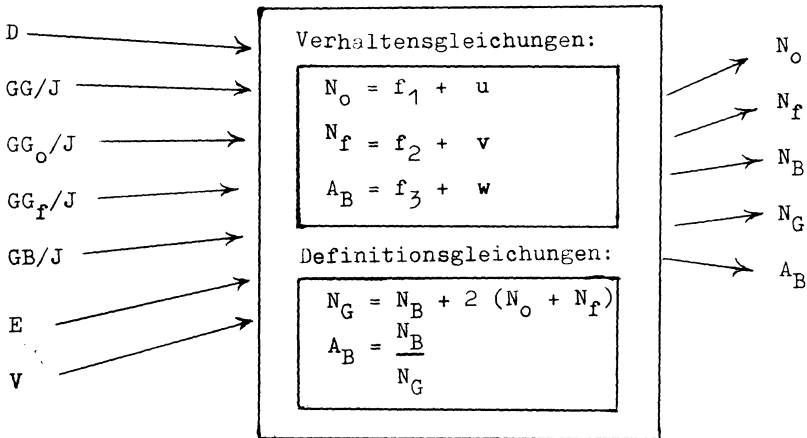
"wenn man so will, der große Müllablageplatz, auf den alle Einflüsse geworfen werden, die man im sauberen, überschaubaren Haushalt des ökonometrischen Modells nicht haben will."

Statistisch gesehen werden die latenten Variablen als Zufallsvariable interpretiert. Ihre Berücksichtigung bewirkt den stochastischen Charakter des ökonometrischen Gleichungssystems und ist Voraussetzung für die statistische Schätzung für die Parameter und damit für die Strukturbestimmung des Modells.

Wie man sich leicht überzeugt, ist die Voraussetzung der eindeutigen Identifizierbarkeit der Struktur in unserem Modell gegeben.

Die Durchrechnung des Modells mit dem gegebenen Datenmaterial lieferte annehmbare Ergebnisse für die Parameter sowie für die Korrelationskoeffizienten, die Durbin-Watson-Werte u. ä. Testgrößen.

Die numerischen Werte der Parameter bestimmen die Struktur des Modells. Kennt man die Werte der exogenen Variablen für den Prognosezeitraum, so lassen sich Prognosewerte für die endogenen Variablen bestimmen.



Die Parameter werden geschätzt aufgrund der in der Vergangenheit beobachteten empirischen Datenreihen für die endogenen und exogenen Variablen. Je länger der Zeitraum ist, der von den Beobachtungsreihen abgedeckt wird, d. h. je mehr Daten zur Verfügung stehen, um so besser wird die Schätzgenauigkeit sein. Nur: Je größer der Beobachtungszeitraum ist, um so problematischer wird die grundlegende Annahme der Strukturkonstanz, von MENGES und DIEHL auch als "Stabilitätsproblem der Ökonometrie" gekennzeichnet:

Im Relevanzzeitraum, d. h. sowohl im Beobachtungszeitraum als auch im Prognosezeitraum treten keine neuen Einflußfaktoren hinzu und es scheiden auch keine durch Bedeutungsminderung aus; die Intensität ihres Einflusses bleibt konstant.

Möglichkeiten zur Überwindung des Dilemmas einer konstanten ökonometrischen Struktur sind:

eine differenziertere Modellspezifikation und der Versuch, die Zahlen der Vergangenheit, die der Prognose zugrunde liegen, durch unterschiedliche Gewichte zu aktualisieren.

Als vielversprechendster Ausweg aus der Sackgasse bietet sich die Simulationstechnik an.

Stochastische Prozesse und makroökonomische Konjunkturtheorie
 von B. Ships, Bochum

Neben den zahlreichen deterministischen¹⁾ finden sich in der Literatur auch einige stochastische Konjunkturmodelle²⁾. Bei solchen stochastischen Konjunkturmodellen wird in der Regel die Konjunktur als zufällige Abweichung von der langfristigen Entwicklung aufgefaßt. Die Grundkonzeption eines derartigen Konjunkturmodells sieht dann etwa folgendermaßen aus³⁾:

Das Volkseinkommen wachse langfristig mit einer konstanten Rate, die durch die Bevölkerungsentwicklung und den technischen Fortschritt bestimmt sei. Es wird angenommen, daß sich Konsumenten und Investoren in ihren Dispositionen diesem Wachstumspfad anpassen. Der von der langfristigen Entwicklung des Konsums kurzfristig abweichende Konsum sei eine Funktion der durchschnittlichen Einkommenserwartungen aller Konsumenten, und diese Erwartungen wiederum seien Realisationen von Zufallsvariablen. Auch wenn man gewisse Abhängigkeiten dieser Zufallsvariablen unterstellt, ist diese Abweichung nach dem Grenzwertsatz von Bernstein eine Realisation einer asymptotisch normalverteilten Zufallsvariablen⁴⁾.

1) Vgl. z.B. R.V. Clemence, A.H. Hansen, Readings in Business Cycles and National Income, New York 1953

2) Vgl. etwa W. Krelle, Grundlinien einer stochastischen Konjunkturtheorie, in: Zeitschrift für die gesamte Staatswissenschaft 115 (1959), S. 472 ff.
 W. Stier, Makroökonomische Anwendungen von Operations Research-Methoden, Meisenheim 1970

3) Vgl. z.B. W. Stier, Makroökonomische..., a.a.O.

4) Vgl. S. Bernstein, Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes, in: Math. Annalen 97 (1927), S. 14 ff.

Bei entsprechenden Überlegungen zu den von der langfristigen Entwicklung der Investitionen abweichenden Investitionen, ergibt sich dann für das vom langfristigen Wachstumspfad abweichende Volkseinkommen eine Summe von zwei zeitlich invariant verteilten Zufallsvariablen, so daß sich die konjunkturelle Entwicklung als eine Realisation eines stationären, stochastischen Prozesses mit zufälliger Amplitude und Phase darstellen läßt.

So sehr nun ein solches Modell dem betrachteten Phänomen entspricht, so wenig hält es in dieser Form methodologischer Kritik stand. Solche Modelle sind ohne empirischen Gehalt. Trotzdem ist die Theorie stationärer stochastischer Prozesse für die Formulierung von Konjunkturtheorien von Bedeutung. Denn die darauf aufbauende Spektralanalyse kann ein Hilfsmittel für die Konjunkturdiagnose und -prognose sein und Hinweise für die Konstruktion und Überprüfung ökonometrischer Modelle geben. Mit Hilfe der Spektralanalyse kann etwa untersucht werden, ob eine Approximation der konjunkturellen Entwicklung durch ein lineares Modell angebracht ist oder nicht und welche lead-lag-Beziehungen zwischen den einzelnen Größen bestehen. Diese Art von Analyse gestattet auch eine Untersuchung der Frage, ob und inwieweit sich bestimmte Größen lang- oder kurzfristig beeinflussen, und sie ermöglicht es, ökonometrische Modelle durch den Vergleich von Spektren der mit Hilfe eines Modells simulierten Zeitreihen und der beobachteten Zeitreihen, zu testen¹⁾.

Bei der Spektralanalyse wird eine Zeitreihe als eine Realisation eines diskreten, kovarianz-stationären stochastischen Prozesses aufgefaßt. Ein solcher Prozeß läßt sich im Zeitbereich durch die sogenannte Autokovarianzfunktion und im Frequenzbereich

1) Vgl. dazu G.M. Jenkins, A Survey of Spectral Analysis, in: Applied Statistics 14 (1965) S. 48 ff., T.H. Naylor, K. Wertz, T.H. Wonnacott, Spectral Analysis of Data Generated by Simulation Experiments with Econometric Models, in: Econometrica 37 (1969) S. 333 ff.

durch das Spektrum beschreiben. Autokovarianzfunktion und Spektrum haben zwar den gleichen Informationsgehalt, das Spektrum bietet jedoch bessere Interpretationsmöglichkeiten. Die Spektralanalyse kann nämlich als eine Weiterentwicklung der klassischen Varianzanalyse aufgefaßt werden. Die Gesamtvarianz des stochastischen Prozesses setzt sich aus den Teilvarianzen der einzelnen Frequenzen zusammen und aus der Verteilung der Gesamtvarianz ergeben sich dann Informationen über den Beitrag der einzelnen Frequenzen zur Gesamtvarianz und damit Informationen über die relative Bedeutung zyklischer Bewegungen verschiedener Frequenzen in bezug auf die Entwicklung des gesamten Prozesses ¹⁾ Diese Art von Information liefert in einfacher Weise das Spektrum.

Es gibt nun direkte und indirekte Methoden zur Schätzung eines Spektrums. Bei den indirekten Methoden wird das Spektrum über die Autokovarianzfunktion geschätzt, während bei den direkten Methoden aus den Zeitreihenwerten das sogenannte Periodogramm ermittelt und als Schätzfunktion für das Spektrum verwendet wird. In beiden Fällen erhält man konsistente Schätzfunktionen jedoch erst durch die Einführung von bestimmten Gewichtsfunktionen ²⁾.

Bei einer Kreuzspektralanalyse werden zwei Zeitreihen als eine Realisation eines bivariaten, diskreten, stationären stochastischen Prozesses betrachtet. Kohärenz-, Phasen- und Transferfunktion geben dann Aufschluß über die Beziehungen zwischen den beiden Zeitreihen. Die Kohärenzfunktion ist ein Maß für den linearen Zusammenhang zwischen den korrespondierenden Frequenzkomponenten der beiden Prozesse. Die Kohärenz entspricht also dem einfachen, linearen Determinationskoeffizienten und ist entsprechend normiert. Die Phasenfunktion ergibt ein Maß für die Phasendifferenz entsprechender Frequenzen und die Transferfunktion gibt an, wie sich Veränderungen der einen Größe auf die andere auswirken.

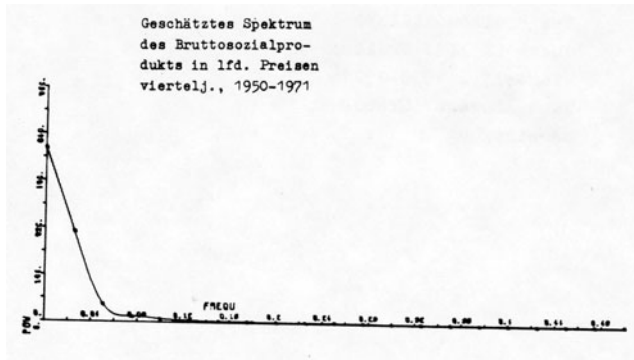
1) Vgl. J.W. Tukey, Discussion, Emphasizing the Connection between Analysis of Variance and Spectrum Analysis, in: *Technometrics* 3 (1961), S. 191 ff.

2) Vgl. C. Bingham, M.D. Godfrey, J.W. Tukey, Modern Techniques of Power Spectrum Estimation, in *IEEE Transactions on Audio and Electroacoustics* 15 (1967) S. 56 ff.

Bei der praktischen Anwendung der Methoden der Spektralanalyse gibt es jedoch noch eine Reihe von ungelösten Problemen. Noch weitgehend ungeklärt ist z.B. die Frage der Auswirkung alternativer Gewichtsfunktionen bei der Schätzung des Spektrums, wenn nur relativ kurze Zeitreihen zur Verfügung stehen und z.B. die Behandlung der Nichtstationarität von ökonomischen Zeitreihen. Die Durchführung einer Spektralanalyse ist deshalb ein jedesmal individuell anzupackendes Problem, wobei man sich stets vergewissern sollte, daß bestimmte Ergebnisse nicht nur auf ein bestimmtes Vorgehen zurückzuführen sind. Ohne auf diese Probleme hier näher einzugehen, soll nun, unter Verzicht auf die Technik, kurz über die Ergebnisse einer Spektralanalyse der wichtigsten, verfügbaren makroökonomischen Zeitreihen für die BRD in dem Untersuchungszeitraum von 1950 - 1971 berichtet werden.

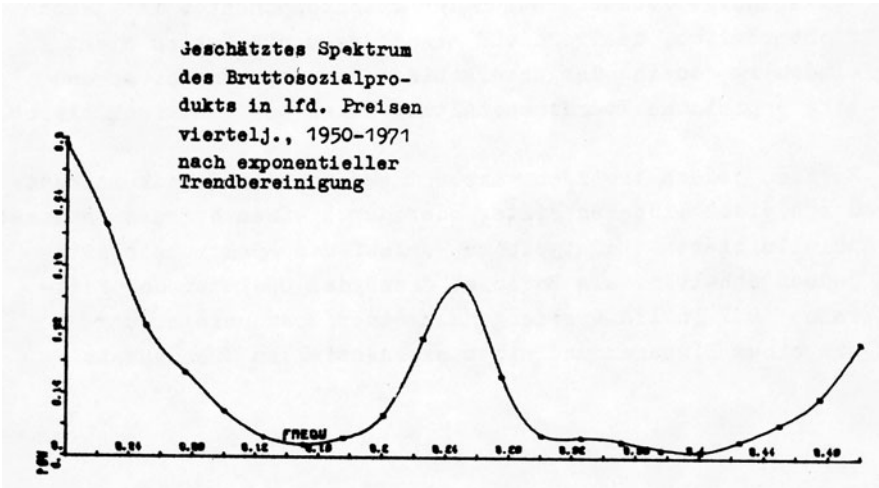
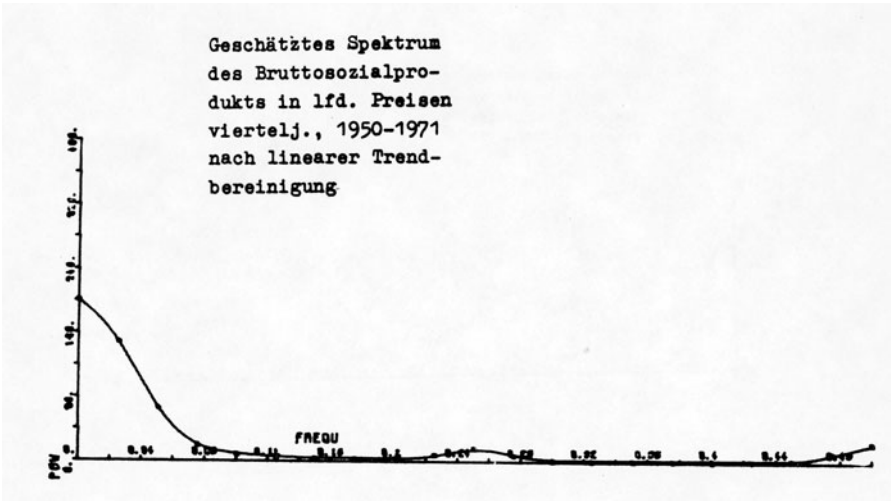
Die Autospektren praktisch aller untersuchten Zeitreihen zeigen den schon von Granger für ökonomische Zeitreihen festgestellten typischen Verlauf¹⁾. Sie zeigen eine Konzentration spektraler Masse nur im Niederfrequenzbereich sowie in den saisonalen Frequenzbändern. Bemerkenswert ist, daß sich kein, oder nur ein schwach ausgeprägter, einem Konjunkturzyklus entsprechender, mittelfristiger Zyklus feststellen läßt. Als Beispiel dafür möge das Spektrum der Zeitreihe 'BSP in lfd. Preisen' dienen.

¹⁾ Vgl. C.W.J. Granger, The Typical Spectral Shape of an Economic Variable, in: *Econometrica* 34 (1966), S. 150 ff.



Eine Ursache dafür, daß sich der aus deskriptiven Analysen einzelner Reihen bekannte Zyklus mit einer Periode von ca. 60 Monaten im Spektrum dieser Reihen nicht feststellen läßt, ist in der Konzentration spektraler Masse in dem Frequenzband um Null zu sehen. Für diese Konzentration ist die Trendkomponente in der Entwicklung dieser Größen verantwortlich. Eine vollständige Identifikation der Trendkomponenten ist jedoch nicht möglich, da Trend und niederfrequente Zyklen nicht eindeutig voneinander unterschieden werden können, so daß eine empirische Trendausschaltung immer problematisch bleibt.

Es kann jedoch trotzdem versucht werden, die Trendkomponente durch einen linearen Filter oder durch einen Regressionsansatz zu eliminieren. Der typische Verlauf des Spektrums bleibt jedoch erhalten. Als Beispiel diene das Spektrum der Zeitreihe 'BSP in lfd.Preisen' nach einer Trendbereinigung mit einem linearen und einem exponentiellen Trendansatz.

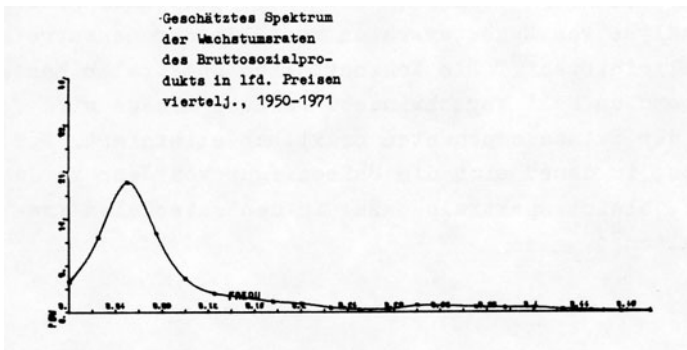


Zeigen die Autospektren also den von Granger festgestellten typischen Verlauf, so decken sich die Ergebnisse der Kreuzspektralanalysen mit den von Godfrey für die USA gemachten Feststellungen¹⁾. Instabile Kohärenz- und Phasenfunktionen und unregelmäßig verlaufende Transferfunktionen legen den Schluß nahe, daß zwischen diesen makroökonomischen Gesamtgrößen nirgends gesicherte lead-lag-Beziehungen bestehen und daß zwischen den einzelnen Größen keine einfachen, d.h. linearen Abhängigkeiten bestehen.

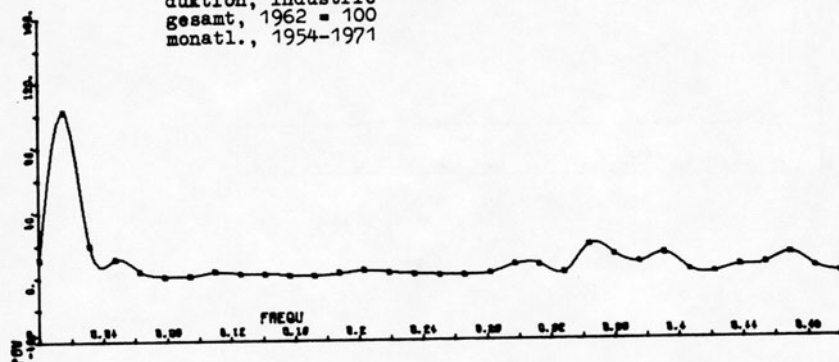
Für diese Resultate könnte jedoch auch die offensichtliche Nicht-Stationarität der betrachteten Zeitreihen verantwortlich sein²⁾. Das Problem der Nicht-Stationarität läßt sich jedoch durch die Analyse von Wachstumsraten anstelle der Gesamtreihen weitgehend eliminieren³⁾. Die Konzentration spektraler Masse im Frequenzband um Null verschwindet. Darüber hinaus wird der Einfluß der Saisonkomponenten praktisch eliminiert. Nur in den Fällen, in denen sich die Saisonfigur von Jahr zu Jahr stark ändert, bleibt spektrale Masse in den saisonalen Frequenzen erhalten⁴⁾.

-
- 1) Vgl. M.D. Godfrey, *Frequency Methods to Economic Analysis*, London 1962
 - 2) Stationär bedeutet hier nicht 'keine Veränderung im Zeitablauf' sondern ein bestimmtes Autokorrelationsverhalten des Prozesses.
 - 3) Unter Wachstumsraten werden dabei die prozentualen Veränderungen einer Größe gegenüber ihrem Wert im Vorjahr, bezogen auf diesen Wert verstanden.
 - 4) Ein Beispiel dafür ist etwa die Zeitreihe 'Wachstumsraten des Indexes der industriellen Nettoproduktion in der Bauindustrie'.

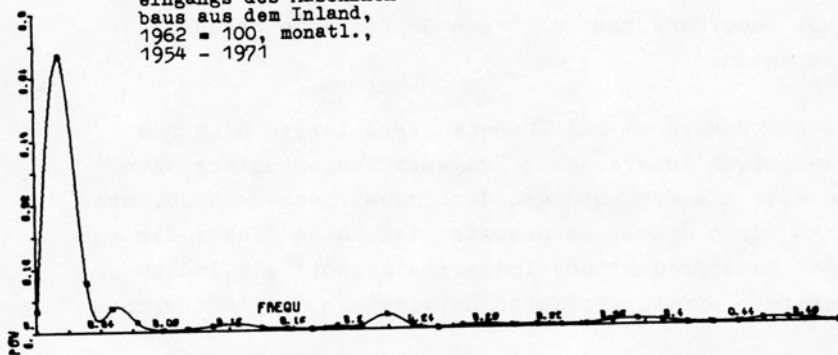
Die Autospektren der Wachstumsraten zeigen dann ein überraschend einheitliches Bild. Fast alle Reihen zeigen im Spektrum eine Konzentration spektraler Masse praktisch nur in dem Frequenzband, das einem Zyklus mit einer Periode von ungefähr 60 Monaten entspricht. Zur Illustration mögen die folgenden Beispiele dienen.

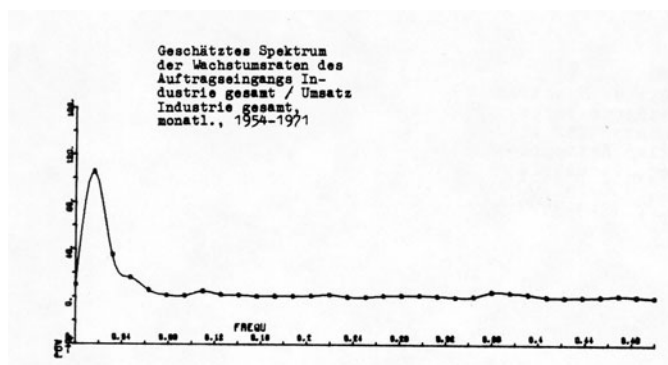


Geschätztes Spektrum
der Wachstumsraten
des Indexes der indu-
striellen Nettopro-
duktion, Industrie
gesamt, 1962 = 100
monatl., 1954-1971



Geschätztes Spektrum
der Wachstumsraten des
Indexes des Auftrags-
einkangs des Maschinen-
baus aus dem Inland,
1962 = 100, monatl.,
1954 - 1971





Aufgrund dieser Spektren darf angenommen werden, daß die übrigen Frequenzbänder für die zeitliche Entwicklung der betrachteten Größen praktisch keine Bedeutung haben, d.h. die gesamte Variation der Prozesse wird weitgehend durch dieses, einem Zyklus mit einer Periode von 60 Monaten entsprechende Frequenzband erklärt.

Da die Entwicklung fast aller Reihen also überwiegend durch diesen Zyklus von 60 Monaten bestimmt ist, kann sich die Kohärenz- und Phasenanalyse auf das entsprechende Frequenzband beschränken.¹⁾

Mit Hilfe der Kohärenz- und Phasenanalyse lassen sich nun einige ökonomisch interessante Fragestellungen untersuchen. Nimmt man etwa die Schwankungen der Industrieproduktion, hier ausgedrückt durch die Wachstumsraten der Reihe 'Index der industriellen Nettoproduktion, Industrie gesamt' als Indikator der allgemeinen wirtschaftlichen Entwicklung, so läßt sich

1) Eine Analyse der Phasenbeziehungen ist dabei nur zwischen solchen Frequenzbändern sinnvoll, zwischen denen eine hohe Kohärenz besteht. Unter einer hohen Kohärenz wird dabei hier ein Wert zwischen 0,7 und 1,0 verstanden.

z.B. untersuchen, ob es andere Reihen gibt, deren Wachstumsraten für die so definierte allgemeine wirtschaftliche Entwicklung als 'leading-', 'coinciding-' oder 'lagging-indicators' angesehen werden können und wie groß eine eventuelle Phasenverschiebung dann tatsächlich ist. Eine solche Untersuchung ist dann die geeignete Basis für die Konstruktion sogenannter Diffusionsindizes. Ein Beispiel dafür ist etwa der Konjunkturindikator des Sachverständigenrates¹⁾. Einige der gefundenen Ergebnisse sollen nun dargestellt werden.

Es zeigt sich, daß als Frühindikatoren z.B. die Wachstumsraten der Reihen:

'Index des Auftragseingangs, Industrie gesamt'	(+4 Mon.)
'Index des Auftragseingangs der Investitionsgüterindustrie aus dem Inland'	(+3 Mon.)
'Index des Auftragseingangs des Maschinenbaus aus dem Inland'	(+4 Mon.)
'Auftragseingang/Umsatz, Industrie gesamt'	(+12 Mon.)

in Betracht kommen, nicht dagegen etwa, wie häufig angenommen wird, die Wachstumsraten der Reihe 'Arbeitslose/offene Stellen'.

Es läßt sich ferner die zeitliche Abfolge in der konjunkturellen Entwicklung der einzelnen Produktionsbereiche, d.h. in den einzelnen Branchen, untersuchen. Es läßt sich nachprüfen, ob die vom Sachverständigenrat in seinem Musterzyklus unterstellte defensive Preispolitik²⁾, durch die im konjunkturellen Aufschwung ein Eigendämpfungseffekt entfällt, zutrifft. Es zeigt sich dabei z.B., daß die Wachstumsraten der Reihe 'Index der Erzeugerpreise industrieller Produkte' der allgemeinen wirtschaftlichen Entwicklung erst mit einer Verzögerung von 14 Monaten folgen. Dieses Ergebnis läßt sich durch entsprechende Feststellungen in den einzelnen Produktionsbereichen erhärten.

1) Vgl. 'Ein neues Frühwarnsystem', in: Wirtschaftswoche, Heft 7, 1971.

2) Vgl. dazu die verschiedenen Jahresgutachten des Sachverständigenrates.

Eine weitere, häufig vertretene Hypothese, die Existenz eines auf eine mangelnde Flexibilität der Tariflöhne und -gehälter zurückzuführenden Lohn-lags, läßt sich auf diese Weise ebenfalls bestätigen. So folgen die Wachstumsraten der Reihe 'Bruttolöhne und -gehälter in lfd. Preisen' den Wachstumsraten der Reihe 'BSP in lfd. Preisen' mit einer Verzögerung von 5 Monaten.

Diese wenigen herausgegriffenen Beispiele zeigen bereits, daß das Instrument der Spektralanalyse geeignet ist, Hinweise für die Formulierung und Überprüfung ökonomischer Hypothesen zu geben. So zeigte sich hier, daß für die Konstruktion eines linearen ökonometrischen Modells zur Beschreibung der für die BRD in dem Zeitraum von 1950 - 1971 festgestellten mittelfristigen Wachstumszyklen mit einer Periode von ungefähr 60 Monaten Zeitreihen mit Wachstumsraten verwendet werden sollten.¹⁾

Nun Beispiele für solche in den Wachstumsraten lineare Modelle gibt es bereits. Es sind die Modelle von van der Werf²⁾ und das modifizierte van der Werf Modell von Beckmann/Uebe³⁾. Mit Hilfe der Spektralanalyse läßt sich nun aber auch zeigen, worauf gewisse Schwierigkeiten mit dem van der Werf Modell zurückzuführen sind, nämlich auf die Verwendung preisbereinigter Größen. Das Spektrum der Wachstumsraten der Reihe 'Preisindex des BSP mit Basis 1954' zeigt einen von dem allgemeinen Wachstumszyklus in der Länge abweichenden Zyklus.

-
- 1) Denn bei den Kreuzspektralanalysen der Ursprungswerte ergeben sich ja fast durchweg sehr unregelmäßig verlaufende Transferfunktionen und eine Verteilung der spektralen Masse auf mehrere Frequenzbänder. Dementsprechend müßten die Anhängigkeiten zwischen den betrachteten Reihen in den einzelnen Frequenzbändern gesondert untersucht und berücksichtigt werden.
 - 2) Vgl. D. van der Werf, A Linear Model to Forecast Short Term Movements in the Western German Economy, Amsterdam 1968.
 - 3) M. Beckmann, G. Uebe, Makroökonomische Untersuchungen der Auswirkungen von Steuersystemänderungen, Wiesbaden 1970.

Deshalb darf vermutet werden, daß den Nebenzyklen im Spektrum der Wachstumsraten der Reihe BSP in Preisen von 1954' keine reale Bedeutung zukommt, sondern lediglich auf die Deflationierung zurückzuführen sind.

Ähnliches gilt auch für Spektren der Wachstumsraten anderer preisbereinigter Reihen. Bei der Verwendung dieser Reihen sind also dann die Beziehungen zwischen mehreren Frequenzbändern zu analysieren. Die geschätzten Transferfunktionen deuten jedoch dann darauf hin, daß zwischen den Wachstumsraten der preisbereinigten Reihen keine einfachen, d.h. linearen Beziehungen bestehen. Dieses Resultat läßt sich durch einen weiteren Test stützen. In einem linearen System sind die Parameter unabhängig von den einzelnen Frequenzen. Werden also etwa einzelne Frequenzen ausgefiltert, dann dürfen sich die Parameter nicht signifikant verändern¹⁾. Genau das ist aber hier der Fall.

1) Vgl. M.D. Godfrey, Frequency Methods ... , a.a.O.

Dynamische Aspekte der Aktivitätsanalyse
 von R. Henn und O. Opitz, Karlsruhe und Innsbruck

1. PROBLEMSTELLUNG

In der statischen Aktivitätsanalyse interessiert man sich dafür, aus einer Menge von Produktionsprozessen effiziente zu charakterisieren. In der dynamischen Theorie sucht man nach endlichen bzw. unendlichen Prozeßfolgen; der Begriff der Effizienz ist dabei in geeigneter Form zu übertragen. Das Hauptproblem liegt hier in der Interdependenz je zweier aufeinanderfolgender Glieder der Prozeßfolge, oder genauer, Effizienzprobleme sind in der dynamischen Theorie eng verknüpft mit Entscheidungen, in welcher Höhe der Output einer Periode t als Input der Periode $t + 1$ einzusetzen ist.

In der statischen Theorie ist jede Funktion zur Bewertung zulässiger Prozesse eine monotone Funktion der Produktivität einer Technologie. Bezeichnet man in der dynamischen Theorie die Differenz $y_{t-1} - x_t$ des Outputs in Periode $t - 1$ abzüglich des Inputs in Periode t als Konsumvektor, so wird eine Bewertungsfunktion im dynamischen Sinn durch die Produktivität einzelner Prozesse, sowie vor allem durch Konsumvektoren bestimmt.

Für die folgenden Überlegungen wird ein diskreter, endlicher oder unendlicher Planungszeitraum

$$(1.1) \quad Z = \{1, 2, 3, \dots\}$$

angenommen.

Bezeichnet man mit $x_t \in \mathbb{R}_+^n$ den Input, mit $y_t \in \mathbb{R}_+^n$ den Output zum Zeitpunkt $t \in Z$, ferner mit T_t die Technologie als die Menge der zum Zeitpunkt t zugelassenen Produktionsprozesse (x_t, y_t) , so gilt

$$(1.2) \quad (x_t, y_t) \in T_t \subseteq \mathbb{R}_+^{2n} \quad (t \in Z).$$

Die Aussage $(x_t, y_t) \in T_t$ besagt, im Zeitpunkt t ist mit Hilfe des Inputs x_t ein Output y_t realisierbar. Durch eine Folge $\{(x_t, y_t)\}$ von Produktionsprozessen mit $(x_t, y_t) \in T_t$ werden damit

alle zulässigen Pfade über die Zeit beschrieben.

Folgende generelle Voraussetzungen seien erfüllt:

$$(1.3) \quad T_t = T \quad (t \in \mathbb{Z}); \quad T \text{ konvex, kompakt;}$$

d.h. die Technologie sei über die Zeit konstant und ein konvexer, kompakter Teilraum des \mathbb{R}_+^{2n} .

$$(1.4) \quad (0,0) \in T$$

$$(1.5) \quad \exists (x,y) \in T \text{ mit } x < y.$$

Für die Menge Y_0 der Anfangsvektoren gelte

$$(1.6) \quad Y_0 \cap \mathbb{R}_{++}^{n1) \neq \emptyset \quad \text{und } Y_0 \text{ kompakt.}$$

DEFINITION 1:

Eine Folge $\{(x_t, y_t)\}$ von Prozessen mit $(x_t, y_t) \in T$, $x_t \leq y_{t-1}$ ($t \in \mathbb{Z}$), $y_0 \in Y_0$ heißt endliches bzw. unendliches Programm.

Den Begriff der Effizienz überträgt man aus der statischen Theorie,

DEFINITION 2:

a) Ein Prozeß $(\bar{x}, \bar{y}) \in T$ heißt effizient, wenn es kein $(x, y) \in T$ gibt mit $(-x, y) \geq (-\bar{x}, \bar{y})$ und $(-x, y) \neq (-\bar{x}, \bar{y})$.

b) Ein Programm $\{(\bar{x}_t, \bar{y}_t)\}$ heißt effizient, wenn es kein anderes zulässiges Programm $\{(x_t, y_t)\}$ gibt mit

$$(-x_t, y_t) \geq (-\bar{x}_t, \bar{y}_t) \quad \text{und} \quad (-x_t, y_t) \neq (-\bar{x}_t, \bar{y}_t) \quad \text{für alle } t \in \mathbb{Z}.$$

Zwischen den beiden Effizienzbegriffen hat man definitionsgemäß den Zusammenhang:

BEMERKUNG 1:

Das Programm $\{(x_t, y_t)\}$ ist genau dann effizient, wenn jeder

1) $\mathbb{R}_{++}^n := \{z \in \mathbb{R}^n : z > 0\}$

Prozeß (x_t, y_t) ($t \in \mathbb{Z}$), wobei $x_t \leq y_{t-1}$, (im statischen Sinne) effizient ist.

Eine zentrale Rolle in der dynamischen Theorie spielen sogenannte stationäre Programme.

DEFINITION 3:

Ein Programm $\{(x_t, y_t)\}$ heißt stationär, wenn für alle $t \in \mathbb{Z}$ gilt $x_t = x$, $y_t = y$, d.h., wenn $\{(x_t, y_t)\}$ eine konstante Folge $\{(x, y)\}$ mit $x \leq y$ darstellt.

Man kann in der statischen wie in der dynamischen Theorie effiziente Prozesse bzw. Programme charakterisieren, indem man Bewertungen vornimmt (vgl. in [4], [6], [9]).

Effiziente Prozesse bzw. Programme ergeben sich dann im Falle eines endlichen Planungszeitraums als Extremalstellen von geeigneten Optimierungsproblemen.

In den folgenden Überlegungen wird eine Bewertungs- oder Nutzenfunktion $u : T \rightarrow \mathbb{R}^1$ zugrundegelegt, die folgende Kriterien erfüllen soll:

$$(1.7) \quad u : T \rightarrow \mathbb{R}^1 \text{ stetig, konkav}$$

$$(1.8) \quad \sup_{(x,y) \in T} \frac{u(x,y) - u(x',y')}{|(x,y) - (x',y')|} < \infty \text{ für alle } (x',y') \in T$$

Die Bedingung (1.8) läßt sich durch eine Regularitätsbedingung an T ersetzen (vgl. [3], Seite 8).

Die Problemstellung liegt nun in der Lösung eines Optimierungsproblems

$$(1.9) \quad \max_t \sum_t u(x_t, y_t) \quad \text{mit} \quad (x_t, y_t) \in T.$$

Im Falle eines endlichen Planungszeitraums ist die Existenz einer Lösung des Problems (1.9) wegen (1.3) und (1.7) immer

gesichert. Als Lösungsverfahren bieten sich Methoden der dynamischen Optimierung an.

Im Falle eines unendlichen Planungszeitraums wird die angegebene Reihe $\sum_t u(x_t, y_t)$ im allgemeinen divergieren. Man behilft sich dann mit Optimalitätskriterien, die etwa auf Ramsey [10], v. Weizsäcker [13] oder Gale [6] zurückgehen.

DEFINITION 4:

Ein Programm $\{(\bar{x}_t, \bar{y}_t)\}$ dominiert (streng) ein Programm $\{(x_t, y_t)\}$, oder kurz $\{(\bar{x}_t, \bar{y}_t)\} \triangleright (\triangleright) \{(x_t, y_t)\}$

wenn $\lim_{N \rightarrow \infty} \sum_{t=1}^N (u(\bar{x}_t, \bar{y}_t) - u(x_t, y_t)) \geq 0$ ¹⁾

$(\exists N_0 : \sum_{t=1}^N (u(\bar{x}_t, \bar{y}_t) - u(x_t, y_t)) > 0 \text{ für alle } N \geq N_0$

Mit Hilfe des vorliegenden Dominanzbegriffes werden verschiedene Optimalitätskriterien angegeben.

DEFINITION 5:

- a) Ein Programm $\{(\bar{x}_t, \bar{y}_t)\}$ heißt (streng) optimal, wenn für alle Programme $\{(x_t, y_t)\}$ gilt: $\{(\bar{x}_t, \bar{y}_t)\} \triangleright (\triangleright) \{(x_t, y_t)\}$.
- b) Ein Programm $\{(\bar{x}_t, \bar{y}_t)\}$ heißt (schwach) maximal, wenn kein anderes zulässiges Programm $\{(x_t, y_t)\}$ existiert, so daß gilt $\{(x_t, y_t)\} \triangleright (\triangleright) \{(\bar{x}_t, \bar{y}_t)\}$.

Optimal bzw. streng optimal heißt demnach ein Programm genau dann, wenn es alle anderen Programme im Sinne von Definition 4 dominiert bzw. streng dominiert. (vgl. [3], Seite 2).

¹⁾ lim bedeutet „limes inferior“, lim analog dazu „limes superior“.

Dagegen korrespondiert der Begriff der Maximalität mit dem der Effizienz (vgl. Definition 2) : Maximal heißt ein Programm $\{(\bar{x}_t, \bar{y}_t)\}$ genau dann, wenn es kein anderes Programm gibt, welches das Programm $\{(\bar{x}_t, \bar{y}_t)\}$ dominiert.

Zwischen den einzelnen Optimalitätsbegriffen ergibt sich der folgende Zusammenhang:

SATZ 1:

- a) $\{(\bar{x}_t, \bar{y}_t)\}$ streng optimal $\Rightarrow \{(\bar{x}_t, \bar{y}_t)\}$ optimal $\Rightarrow \{(\bar{x}_t, \bar{x}_t)\}$ schwach maximal
 b) $\{(\bar{x}_t, \bar{y}_t)\}$ streng optimal $\Rightarrow \{(\bar{x}_t, \bar{y}_t)\}$ maximal $\Rightarrow \{(\bar{x}_t, \bar{x}_t)\}$ schwach maximal

Beweis:

- a) $\{(\bar{x}_t, \bar{y}_t)\}$ streng optimal $\Leftrightarrow \bigcup_{N_0} \sum_{t=1}^N (u(\bar{x}_t, \bar{y}_t) - u(x_t, y_t)) > 0$ für alle $N \geq N_0$

und alle Programme $\{(x_t, y_t)\}$ mit $(x_t, y_t) \in T$ und $x_t \leq y_{t-1}$

$$\Rightarrow \lim_{N \rightarrow \infty} \sum_{t=1}^N (u(\bar{x}_t, \bar{y}_t) - u(x_t, y_t)) > 0 \Rightarrow \{(\bar{x}_t, \bar{y}_t)\} \text{ optimal.}$$

Angenommen, $\{(\bar{x}_t, \bar{y}_t)\}$ sei nicht schwach maximal, dann gäbe es nach Definition 4 und 5 ein Programm $\{(\tilde{x}_t, \tilde{y}_t)\}$ und

ein $N_1 \in \mathbb{N}$ mit: $\sum_{t=1}^N (u(\tilde{x}_t, \tilde{y}_t) - u(\bar{x}_t, \bar{y}_t)) > 0$ für alle $N \geq N_1$,

woraus folgt $\lim_{N \rightarrow \infty} \sum_{t=1}^N (u(\bar{x}_t, \bar{y}_t) - u(\tilde{x}_t, \tilde{y}_t)) < 0$, im Widerspruch zur Optimalität von $\{(\bar{x}_t, \bar{y}_t)\}$.

- b) Angenommen, $\{(\bar{x}_t, \bar{y}_t)\}$ sei streng optimal, jedoch nicht maximal.

Dann gibt es ein Programm $\{(\hat{x}_t, \hat{y}_t)\}$ mit

$$\lim_{N \rightarrow \infty} \sum_{t=1}^N (u(\hat{x}_t, \hat{y}_t) - u(\bar{x}_t, \bar{y}_t)) \geq 0 \text{ im Widerspruch zur strengen}$$

Optimalität von $\{(\bar{x}_t, \bar{y}_t)\}$, welche impliziert:

$$\exists N_0 : \sum_{t=1}^N (u(\bar{x}_t, \bar{y}_t) - u(\hat{x}_t, \hat{y}_t)) > 0 \text{ für alle } N \geq N_0.$$

Aus der Maximalität folgt trivialerweise die schwache Maximalität.

KOROLLAR:

Die Reihe $\sum_{t=1}^{\infty} u(x_t, y_t)$ sei konvergent für alle Programme $\{(x_t, y_t)\}$.

Dann gilt:

- a) $\{(\bar{x}_t, \bar{y}_t)\}$ streng optimal $\Leftrightarrow \{(\bar{x}_t, \bar{y}_t)\}$ maximal
- b) $\{(\bar{x}_t, \bar{y}_t)\}$ optimal $\Leftrightarrow \{(\bar{x}_t, \bar{y}_t)\}$ schwach maximal.

Beweis:

a) " \Leftarrow " Wenn kein Programm $\{(x_t, y_t)\}$ existiert mit

$$\lim_{N \rightarrow \infty} \sum_{t=1}^N (u(x_t, y_t) - u(\bar{x}_t, \bar{y}_t)) \geq 0, \text{ so folgt wegen } \sum_{t=1}^{\infty} u(x_t, y_t)$$

$$\text{konvergent } \lim_{N \rightarrow \infty} \sum_{t=1}^N (u(x_t, y_t) - u(\bar{x}_t, \bar{y}_t)) = \lim_{N \rightarrow \infty} \sum_{t=1}^N (u(x_t, y_t) - u(\bar{x}_t, \bar{y}_t))$$

$$= \lim_{N \rightarrow \infty} \sum_{t=1}^N (u(x_t, y_t) - u(\bar{x}_t, \bar{y}_t)) < 0 \text{ für alle zulässigen Programme } \{(x_t, y_t)\}.$$

Also gilt:

$$\lim_{N \rightarrow \infty} \sum_{t=1}^N (u(\bar{x}_t, \bar{y}_t) - u(x_t, y_t)) = \lim_{N \rightarrow \infty} \sum_{t=1}^N (u(\bar{x}_t, \bar{y}_t) - u(x_t, y_t)) > 0$$

und damit ist $\{(\bar{x}_t, \bar{y}_t)\}$ wegen der Stetigkeit von u streng optimal.

Die Umkehrung wurde in Satz 1 gezeigt.

b) " \Leftarrow " $\{(\bar{x}_t, \bar{y}_t)\}$ ist schwach maximal, d.h., es gibt kein Programm

$$\{(x_t, y_t)\} \text{ mit } \sum_{t=1}^N (u(x_t, y_t) - u(\bar{x}_t, \bar{y}_t)) > 0 \text{ für } N \geq N_0, \text{ daraus folgt,}$$

wegen $\sum_{t=1}^{\infty} u(x_t, y_t)$ konvergent, $\lim_{N \rightarrow \infty} \sum_{t=1}^N (u(\bar{x}_t, \bar{y}_t) - u(x_t, y_t))$
 $= \lim_{N \rightarrow \infty} \sum_{t=1}^N (u(\bar{x}_t, \bar{y}_t) - u(x_t, y_t)) \geq 0$, also ist $\{(\bar{x}_t, \bar{y}_t)\}$ optimal.

Die Umkehrung wurde in Satz 1 gezeigt.

2. EXISTENZAUSSAGEN

Zum Problem der Existenz von streng optimalen, maximalen und schwach maximalen Programmen sollen zwei Beispiele Aufschluß geben (vgl. dazu [3]).

Beispiel 1:

Gegeben sei die Technologie $T_t := \{(x_t, y_t) = (x_{1t}, x_{2t}, y_{1t}, y_{2t}) \in \mathbb{R}_+^4 : y_{1t} = y_{2t} = c x_{2t}\}$ mit $c > 1$. Man interpretiere die Inputs x_{1t}, x_{2t}

als Konsum bzw. Investition sowie die Outputs $y_{1t} = y_{2t}$

als Sozialprodukt einer Wirtschaft, der Faktor $c > 1$ stellt die Wachstumsrate dar. Auf Grund der Ex-post-Beziehung $y_{1,t-1} = y_{2,t-1} = x_{1t} + x_{2t}$

erfüllt jedes zulässige Programm $\{(x_t, y_t)\}$ mit $x_t = (x_{1t}, x_{2t})$

und $y_t = (y_{1t}, y_{2t})$ und der Anfangsbedingung $y_0 = (1, 1)$ die Bedingungen $(x_{11}, x_{21}) \leq (1, 1)$

und $(x_{1t}, x_{2t}) = (y_{1,t-1}, y_{2,t-1}) = (c x_{2,t-1}, c x_{2,t-1})$ für $t = 2, 3, \dots$

Die auf T erklärte Bewertungsfunktion sei gegeben durch

$$u(x_t, y_t) = \begin{cases} x_{11} = 1 - x_{21} & \text{für } t = 1 \\ x_{1t} = y_{1,t-1} - x_{2t} \\ = c x_{2,t-1} - x_{2t} & \text{für } t = 2, 3, \dots \end{cases}$$

Sei nun $\{(x_t, y_t)\}$ ein beliebiges Programm mit $u(x_{t_0}, y_{t_0}) = x_{1t_0} > 0$

für mindestens ein $t_0 \in \mathbb{N}$ und $\{(\hat{x}_t, \hat{y}_t)\}$ ein Programm mit $\hat{x}_t = x_t$

für alle $t \neq t_0$ (also auch $\hat{y}_t = y_t$ für $t \neq t_0$) und $x_{2t_0} < \hat{x}_{2t_0}$.

Es ist also

$$\begin{aligned}
 u(\hat{x}_t, \hat{y}_t) &= u(x_t, y_t) \text{ für } t \notin \{t_0, t_0 + 1\}. \text{ Für } t_0 \in \{2, 3, \dots\} \\
 \text{erhält man } u(x_{t_0}, y_{t_0}) + u(x_{t_0+1}, y_{t_0+1}) &= x_{1t_0} + x_{1,t_0+1} \\
 &= c x_{2,t_0-1} - x_{2t_0} + c x_{2t_0} - x_{2,t_0+1} \\
 &= c x_{2,t_0-1} - x_{2,t_0+1} + (c-1) x_{2t_0} < c x_{2,t_0-1} - x_{2,t_0+1} + (c-1) \hat{x}_{2t_0} \\
 &= c \hat{x}_{2,t_0-1} - \hat{x}_{2,t_0+1} + (c-1) \hat{x}_{2t_0} = \hat{x}_{1t_0} + \hat{x}_{1,t_0+1} \\
 &= u(\hat{x}_{t_0}, \hat{y}_{t_0}) + u(\hat{x}_{t_0+1}, \hat{y}_{t_0+1}).
 \end{aligned}$$

Für $t_0 = 1$ setze man $1 = c x_{20}$. Dann ist obige Abschätzung gleichfalls richtig.

Daraus folgt:

Zu jedem Programm $\{(x_t, y_t)\}$ mit $u(x_t, y_t) > 0$ für mindestens ein $t \geq 1$ existiert ein Programm $\{(\hat{x}_t, \hat{y}_t)\}$ und ein $N_0 \in \mathbb{N}$ mit:

$$\sum_{t=1}^N (u(\hat{x}_t, \hat{y}_t) - u(x_t, y_t)) > 0 \text{ für alle } N \geq N_0.$$

Das Programm $\{(x_t, y_t)\}$ mit $u(x_t, y_t) = 0$ für alle t ist das bzgl. u schlechteste zulässige Programm. Gemäß Definition 5 existiert in Beispiel 1 kein schwach maximales Programm und damit nach Satz 1 auch kein maximales, optimales oder streng optimales Programm.

Beispiel 2:

Gegeben sei die Technologie T_t des Beispiels 1 mit $c > 1$ und der Anfangsbedingung $y_0 = (1, 1)$. Die Variablen sind wie im Beispiel 1 zu interpretieren, auch hier gilt $y_{1,t-1} = y_{2,t-1} = x_{1t} + x_{2t}$, zusätzlich gelte jedoch $x_{2t} \leq \min(1, y_{2,t-1})$. Die Investition ist also im Vergleich zu Beispiel 1 durch die zusätzliche Nebenbedingung $x_{2t} \leq 1$ nach oben beschränkt. Jedes zulässige Programm $\{(x_t, y_t)\}$ erfüllt die Bedingung

$$(x_{11}, x_{21}) \leq (1, 1) \text{ und } (x_{1t}, x_{2t}) \leq (y_{1t-1}, y_{2t-1}) \quad \text{für } t = 2, 3, \dots$$

Die Bewertungsfunktion sei analog zu Beispiel 1 gegeben durch

$$u(x_t, y_t) = \begin{cases} x_{11} = 1 - x_{21} & \text{für } t = 1 \\ x_{1t} = c \cdot x_{2,t-1} - x_{2t} & \text{für } t = 2, 3, \dots \end{cases}$$

Betrachtet man ein beliebiges Programm $\{(x_t, y_t)\}$ mit $x_{2t} = 1$ u. $x_{2t'} < 1$ für mindestens ein $t' \in \mathbb{N}$ und vergleicht dieses Programm mit $\{(\bar{x}_t, \bar{y}_t)\}$, wobei $\bar{x}_{2t} = \bar{y}_{1t} = \bar{y}_{2t} = 1$ und $\bar{x}_{1t} = c \cdot x_{2,t-1} - 1 = c - 1$ für alle t , nimmt man ferner an, $\{(\bar{x}_t, \bar{y}_t)\} \nVdash \{(x_t, y_t)\}$, dann existiert eine Folge t_0, t_1, t_2, \dots mit $t_0 = \min\{t' : x_{2t'} < 1\}$, so daß für $n = 1, 2, \dots$

$$\text{gilt: } 0 \leq \sum_{t=t_0}^{t_n} (u(x_t, y_t) - (c-1)) .$$

Mit $c \cdot x_{20} = 1$ folgt daraus

$$\begin{aligned} 0 &\leq c \cdot x_{2,t_0-1} - x_{2t_0} - (c-1) + c \cdot x_{2t_0} - x_{2,t_0+1} - (c-1) + \dots \\ &\quad + c \cdot x_{2,t_n-2} - x_{2,t_n-1} - (c-1) + c \cdot x_{2,t_n-1} - x_{2t_n} - (c-1) = \\ &= -c(1-x_{2,t_0-1}) + 1 - x_{2t_0} - c(1-x_{2t_0}) + 1 - x_{2,t_0+1} - \\ &\quad \dots - c(1-x_{2,t_n-1}) + 1 - x_{2t_n} . \end{aligned}$$

Dies ergibt mit $z_t = 1 - x_{2t} \geq 0$ und $d = c - 1 > 0$

$$0 \leq - (1+d) z_{t_0-1} - d \sum_{t=t_0}^{t_n-1} z_t + z_{t_n}$$

$$\text{oder wegen } - (1+d) z_{t_0-1} \leq 0$$

$$z_{t_n} \geq d \sum_{t=t_0}^{t_n-1} z_t \geq d \sum_{i=0}^{n-1} z_{t_i} . \quad (*)$$

$$\text{Ferner gilt: } z_{t_n} \geq (1+d)^{n-1} d z_{t_0} ,$$

$$\text{denn für } n = 1 \text{ folgt direkt aus } (*) \quad z_{t_1} \geq d z_{t_0} .$$

Durch Induktionsschluß erhält man:

$$\begin{aligned} z_{t_{n+1}} &\geq d \sum_{i=0}^n z_{t_i} \geq d (z_{t_0} + d z_{t_0} + d (1+d) z_{t_0} + \dots + d (1+d)^{n-1} z_{t_0}) \\ &= d z_{t_0} (1 + d + d (1+d) + \dots + d (1+d)^{n-1}) \\ &= d z_{t_0} \left(1 + d \left((1+d)^0 + (1+d)^1 + \dots + (1+d)^{n-1} \right) \right) \\ &= d z_{t_0} \left(1 + d \left(\frac{1 - (1+d)^n}{1 - (1+d)} \right) \right) = d z_{t_0} (1+d)^n . \end{aligned}$$

Da aber wegen $d > 0$ $\lim_{n \rightarrow \infty} (1+d)^n = \infty$, existiert ein n mit $z_{t_n} > 1$

und daraus folgt $x_{2 t_n} < 0$, im Widerspruch zur Voraussetzung.

Also war die Annahme falsch, das oben definierte Programm $\{(\bar{x}_t, \bar{y}_t)\}$ dominiert im strengen Sinne jedes andere Programm, $\{(\bar{x}_t, \bar{y}_t)\}$ ist streng optimal, und damit auch optimal, maximal, schwach maximal.

Um Zusammenhänge zwischen den diskutierten Optimalitätskriterien und der "Optimalität" von stationären Programmen (vgl. Definition 3) aufzeigen zu können, formuliert man

DEFINITION 6:

Ein stationäres Programm $\{(x_t, y_t)\}$ mit $x_t = \bar{x}$, $y_t = \bar{y}$

heißt optimales stationäres Programm, wenn gilt

$$(2.1) \quad \max_{\substack{(x,y) \in T \\ x = y}} u(x,y) = u(\bar{x}, \bar{y})$$

Für die Ermittlung eines optimalen stationären Programms sind also nur stationäre Programme zur Konkurrenz zugelassen. Demnach ist ein optimales stationäres Programm nicht notwendig optimal, maximal etc. im Sinne von Definition 5. Wegen T kompakt und u stetig ist jedenfalls die Existenz eines optimalen stationären Programmes gesichert.

Im folgenden werden Existenzaussagen über (schwach) maximale und (streng) optimale Programme formuliert. Ergebnisse hierzu findet man etwa in [1], [3], [7], [11], [12].

DEFINITION 7:

Ein Programm $\{(x_t, y_t)\}$ heißt gut, wenn eine reelle Zahl k existiert mit $k \leq \sum_{t=1}^N (u(x_t, y_t) - u(\bar{x}, \bar{y}))$ für alle $N \in \mathbb{N}$ und $\{(\bar{x}, \bar{y})\}$ ist optimales stationäres Programm.

Ferner findet man in [3] den

SATZ 2 :

Sei $\{(\bar{x}, \bar{y})\}$ ein optimales stationäres Programm; dann gibt es zu jedem zulässigen Programm $\{(x_t, y_t)\}$ eine reelle Zahl K mit $\sum_{t=1}^N (u(x_t, y_t) - u(\bar{x}, \bar{y})) \leq K$ für alle $N \in \mathbb{N}$.

Daraus folgt direkt:

Ein Programm $\{(x_t, y_t)\}$ ist gut oder mit $\{(\bar{x}, \bar{y})\}$ als optimalem stationärem Programm gilt $\lim_{N \rightarrow \infty} \sum_{t=1}^N (u(x_t, y_t) - u(\bar{x}, \bar{y})) = -\infty$

Demnach werden im folgenden ausschließlich gute Programme behandelt.

Dann gilt (vgl. [3], Seite 12)

SATZ 3:

Zu jedem $y_0 > 0$ existiert ein gutes Programm.

Ferner gilt für gute Programme folgende "Turnpike" - Eigenschaft (vgl. [1], Seite 277)

SATZ 4:

Sei $\{(x_t, y_t)\}$ ein gutes Programm und $(\bar{x}_t, \bar{y}_t) = \frac{1}{t} \sum_{n=1}^t (x_n, y_n) \in T$

(wegen T konvex). Dann gilt:

- a) $\lim_{t \rightarrow \infty} u(\bar{x}_t, \bar{y}_t) = u(\bar{x}, \bar{y})$
- b) Wenn $\{(\bar{x}, \bar{y})\}$ eindeutig ist, folgt $\lim_{t \rightarrow \infty} (\bar{x}_t, \bar{y}_t) = (\bar{x}, \bar{y})$.

BROCK zeigt, daß der Nutzen eines Programmes $\{(\bar{x}_t, \bar{y}_t)\}$, welches aus einem guten Programm durch "Durschnittsbildungen" gewonnen wird, gegen den Nutzen eines optimalen stationären Programmes konvergiert.

LEMMA:

- a) Zu jedem Programm $\{(x_t, y_t)\}$ existiert eine nichtnegative Folge $\{\epsilon_t\}$ und ein $q \geq 0$, sodaß gilt

$$(2.2) \quad \sum_{t=1}^N (u(x_t, y_t) - u(\bar{x}, \bar{y})) = q \cdot (y_0 - y_N) - \sum_{t=1}^N \epsilon_t$$

für $N = 1, 2, \dots$

- b) Es existiert ein Programm $\{(\hat{x}_t, \hat{y}_t)\}$, so daß die zugehörige Reihe $\sum_{t=1}^{\infty} \hat{\epsilon}_t$ minimal ist in der Klasse von Programmen mit der Anfangsbedingung y_0 .

Beweis :

- a) Nach dem Satz von Kuhn-Tucker (vgl. [5], Seite 200, 201) ist das optimale stationäre Programm $\{(\bar{x}, \bar{y})\}$ (vgl. Definition 6) Lösung des Problems

$$(2.3) \quad \max_{(x, y) \in T} u(x, y) + q(y - x)$$

und es gilt $q(\bar{y} - \bar{x}) = 0$, $q \geq 0$ ist Lagrange-Multiplikator.

Für beliebiges $(x_t, y_t) \in T$ ist damit

$$(2.4) \quad u(x_t, y_t) + q(y_t - x_t) + \alpha_t = u(\bar{x}, \bar{y}) \text{ mit } \alpha_t \geq 0$$

und summiert über N Glieder des Programmes $\{(x_t, y_t)\}$

$$(2.5) \quad \sum_{t=1}^N (u(x_t, y_t) - u(\bar{x}, \bar{y})) = \sum_{t=1}^N q(x_t - y_t) - \sum_{t=1}^N \alpha_t$$

$$= q(y_0 - y_N) - \sum_{t=1}^N \epsilon_t \quad \text{für } \epsilon_t = \alpha_t - q(x_t - y_{t-1}) \geq 0.$$

- b) vgl. [1] S. 277.

SATZ 5:

Es existiere ein eindeutiges optimales stationäres Programm $\{(\bar{x}, \bar{y})\}$.

Daraus folgt:

- a) Zu jedem $y_0 > 0$ existiert ein schwach maximales Programm, welches von y_0 aus startet.
- b) Ist $\bar{y} \in Y_0$, so ist $\{(\bar{x}, \bar{y})\}$ schwach maximal.

Beweis:

a) Für ein Programm $\{(\hat{x}_t, \hat{y}_t)\}$ mit minimalem $\sum_{t=1}^{\infty} \hat{\varepsilon}_t$ gilt nach (2.2)

$$(2.6) \quad \sum_{t=1}^N u(x_t, y_t) - u(\hat{x}_t, \hat{y}_t) = q (\hat{y}_N - y_N) + \sum_{t=1}^N \hat{\varepsilon}_t - \sum_{t=1}^N \varepsilon_t$$

$$(N = 1, 2, \dots),$$

wobei $\{(x_t, y_t)\}$ ein beliebiges gutes Programm mit der zugehörigen Reihe $\sum_{t=1}^{\infty} \varepsilon_t$ darstellt.

Angenommen $\{(x_t, y_t)\}$ dominiert streng $\{(\hat{x}_t, \hat{y}_t)\}$, so existiert ein $N_0 \in \mathbb{N}$ mit

$$(2.7) \quad \sum_{t=1}^N u(x_t, y_t) - u(\hat{x}_t, \hat{y}_t) > 0 \text{ für alle } N \geq N_0$$

und damit wegen der Minimalität von $\sum_{t=1}^{\infty} \hat{\varepsilon}_t$ auch

$$(2.8) \quad q (\hat{y}_N - y_N) > 0 \text{ für } N \geq N_0.$$

Gemäß Satz 4 gewinnt man aus $\{(x_t, y_t)\}$ bzw. $\{(\hat{x}_t, \hat{y}_t)\}$ die zulässigen Programme $\{(\bar{x}_t, \bar{y}_t)\}$ bzw. $\{(\hat{\bar{x}}_t, \hat{\bar{y}}_t)\}$ durch

$$(\bar{x}_t, \bar{y}_t) = \frac{1}{t} \sum_{n=1}^t (x_n, y_n) \text{ und } (\hat{\bar{x}}_t, \hat{\bar{y}}_t) = \frac{1}{t} \sum_{n=1}^t (\hat{x}_n, \hat{y}_n).$$

Nach (2.8) folgt damit

$$(2.9) \quad \lim_{N \rightarrow \infty} q (\hat{\bar{y}}_N - \bar{y}_N) > 0,$$

während sich aus Satz 4 mit (2.6) ergibt

$$(2.10) \quad 0 = \lim_{N \rightarrow \infty} \sum_{t=1}^N (u(x_t, y_t) - u(\hat{x}_t, \hat{y}_t)) = \lim_{N \rightarrow \infty} q (\hat{\bar{y}}_N - \bar{y}_N).$$

Man erhält einen Widerspruch, also ist $\{(\hat{x}_t, \hat{y}_t)\}$ schwach maximal.

b) Ist $\{(x_t, y_t)\}$ optimal stationär mit $x_t = \bar{x}$, $y_t = \bar{y}$, so gilt (vgl. (2.5))

$$\alpha_t = 0 \text{ und } q (x_t - y_{t-1}) = q (\bar{x} - \bar{y}) = 0, \text{ also } \varepsilon_t = 0.$$

Damit ist $\sum_{t=1}^{\infty} \varepsilon_t = 0$ minimal und nach Beweisschnitt a) ist $\{(\bar{x}, \bar{y})\}$ schwach maximal.

Zur Existenz von optimalen Programmen (vgl. [3], Seite 17)

gilt der

SATZ 6:

u sei streng konkav in (\bar{x}, \bar{y}) und $\{(\bar{x}, \bar{y})\}$ sei optimales, stationäres Programm. Dann existiert unter allen Programmen, welche bei einem $y_0 \in Y_0$ mit $y_0 > 0$ starten, ein optimales Programm $\{(x_t, y_t)\}$.

Beweis:

Sei $\{(\hat{x}_t, \hat{y}_t)\}$ ein Programm zu minimalem $\sum_{t=1}^{\infty} \hat{\epsilon}_t$. Für ein gutes

Programm $\{(x_t, y_t)\}$ gilt nach Definition 7

$$(2.11) \quad k \leq \sum_{t=1}^N (u(x_t, y_t) - u(\bar{x}, \bar{y})),$$

wobei $\{(\bar{x}, \bar{y})\}$ optimales stationäres Programm ist, und damit

nach (2.5)

$$(2.12) \quad k \leq q(y_0 - y_N) - \sum_{t=1}^N \epsilon_t \quad \text{für alle } N = 1, 2, \dots$$

oder für $q(y_0 - y_N) \leq k_0$ ($N = 1, 2, \dots$) die Abschätzung

$$(2.13) \quad \sum_{t=1}^N \epsilon_t = k_0 - k \quad (N = 1, 2, \dots)$$

Daraus folgt:

$$(2.14) \quad \lim_{t \rightarrow \infty} \epsilon_t = 0$$

und wegen $\alpha_t \leq \epsilon_t$ für alle t (vgl. (2.5)) auch

$$(2.15) \quad \lim_{t \rightarrow \infty} \alpha_t = 0.$$

Sei (\tilde{x}, \tilde{y}) ein beliebiger Limespunkt des Programms $\{(x_t, y_t)\}$.

Dann gilt wegen (2.3), (2.4), (2.15)

$$(2.16) \quad u(\bar{x}, \bar{y}) + q(\bar{y} - \bar{x}) = u(\tilde{x}, \tilde{y}) + q(\tilde{y} - \tilde{x}) = \max_{(x, y) \in T} (u(x, y) + q(y - x))$$

mit $q(\bar{y} - \bar{x}) = 0$. Da u in (\bar{x}, \bar{y}) streng konkav ist, gilt

$$(\tilde{x}, \tilde{y}) = (\bar{x}, \bar{y}). \text{ Daraus folgt } \lim_{t \rightarrow \infty} (x_t, y_t) = (\bar{x}, \bar{y}).$$

Analog gilt für das Programm $\{(\hat{x}_t, \hat{y}_t)\}$ mit minimalem $\sum_{t=1}^{\infty} \epsilon_t$ die

Aussage $\lim_{t \rightarrow \infty} (\hat{x}_t, \hat{y}_t) = (\bar{x}, \bar{y})$, da $\{(\hat{x}_t, \hat{y}_t)\}$ ein gutes Programm ist.

Dann folgt aus (2.6) durch Limesübergang

$$\begin{aligned} \lim_{N \rightarrow \infty} \sum_{t=1}^N (u(\hat{x}_t, \hat{y}_t) - u(x_t, y_t)) &= \lim_{N \rightarrow \infty} q(y_N - \hat{y}_N) + \sum_{t=1}^{\infty} \epsilon_t - \sum_{t=1}^{\infty} \hat{\epsilon}_t \\ &= \sum_{t=1}^{\infty} \epsilon_t - \sum_{t=1}^{\infty} \hat{\epsilon}_t \geq 0. \text{ Also ist } \{(\hat{x}_t, \hat{y}_t)\} \text{ optimal.} \end{aligned}$$

BEMERKUNG:

Ist u überall streng konkav, so existiert genau ein streng optimales Programm.

Zum Beweis vgl. [3], Seite 18.

Beispiel 3:

Gegeben sei die Technologie $T := \{(x, y) \in \mathbb{R}_+^2 : x \leq 2y \leq 4x, x+y \leq 3\}$

und die Bewertungsfunktion $u(x_t, y_t) = 2x_t - y_t$.

Bestimmung eines optimalen stationären Programms:

$$\begin{aligned} \max_{(x,y) \in T} (2x - y) &= 2 \cdot 3/2 - 3/2 = 3/2 \\ x &\leq y \quad \text{für } \bar{x} = \bar{y} = 3/2 \end{aligned}$$

Zulässig ist ferner das Programm

$$\{(x_t^1, y_t^1)\} = \{(3/2, 1), (1, 2), (2, 1), (1, 2), \dots\}$$

Es gilt

$$\begin{aligned} \lim_{N \rightarrow \infty} \sum_{t=1}^N (u(x_t, y_t) - u(\bar{x}, \bar{y})) &= -1 \\ \lim_{N \rightarrow \infty} \sum_{t=1}^N (u(x_t, y_t) - u(\bar{x}, \bar{y})) &= 1/2 \end{aligned}$$

Da nach Satz 5 b) das optimale stationäre Programm $\{(\bar{x}, \bar{y})\}$ mit $\bar{x} = \bar{y} = 3/2$ für $y_0 = 3/2$ schwach maximal ist, ist auch das Programm $\{(x_t^1, y_t^1)\}$ schwach maximal. Beide Programme sind auch maximal, jedoch nicht optimal.

3. EXISTENZAUSSAGEN BEI DISKONTIERUNG

Diskontiert man die Nutzen späterer Perioden auf den Anfangszeitpunkt, so bezeichnet man mit

$$(3.1) \quad v \{(x_t, y_t)\} = \sum_{t=1}^{\infty} \delta^{t-1} u(x_t, y_t),$$

wobei $\delta \in \langle 0, 1 \rangle$ einen Diskontfaktor bezeichnet, den Wert eines Programms. Da T kompakt und u stetig ist, konvergiert die angegebene unendliche Reihe. In diesem Fall ist nach Korollar zu Satz 1 die Maximalität eines Programmes gleichwertig mit der strengen Optimalität und die schwache Maximalität gleichwertig mit der Optimalität. Dann ist ein Programm $\{(\bar{x}_t, \bar{y}_t)\}$ genau dann optimal bzw. schwach maximal, wenn für alle Programme $\{(x_t, y_t)\}$ gilt $\sum_{t=1}^{\infty} (u(\bar{x}_t, \bar{y}_t) - u(x_t, y_t)) \geq 0$ und

genau dann streng optimal bzw. maximal, wenn für alle Programme $\{(x_t, y_t)\}$ gilt $\sum_{t=1}^{\infty} u(\bar{x}_t, \bar{y}_t) - u(x_t, y_t) > 0$.

SATZ 7:

- a) Zu jedem $y_0 \in Y_0$ existiert ein optimales Programm $\{(\bar{x}_t, \bar{y}_t)\}$.
 Ist u streng konkav, so ist $\{(\bar{x}_t, \bar{y}_t)\}$ eindeutig.
- b) Es existiert ein stationäres Programm, welches optimal ist.

Beweis:

a) Man definiert die Funktion

$$(3.2) \quad w(y_0) = \sup \{v\{(x_t, y_t)\} : \{(x_t, y_t)\} \text{ Programm mit Start in } y_0\}.$$

Dann gilt für jedes beliebige Programm $\{(x_t, y_t)\}$ die Abschätzung

$$v\{(x_t, y_t)\} \leq u(x_1, y_1) + \delta w(y_1) \leq u(x', y') + \delta w(y')$$

wobei $(x', y') \in T$, $x' \leq y_0$ existiert, da die Funktion $u + \delta w$ stetig ist.

Es existiert also ein $\varepsilon \geq 0$ mit

$$(3.3) \quad 0 \leq \varepsilon = u(x', y') + \delta w(y') - w(y_0).$$

Angenommen, es gelte $\varepsilon > 0$.

Dann existiert wegen der Stetigkeit von u ein Programm $\{(x''_t, y''_t)\}$ von y' aus, so daß $\{(\tilde{x}_t, \tilde{y}_t)\}$ mit $(\tilde{x}_1, \tilde{y}_1) = (x', y')$ und $(\tilde{x}_t, \tilde{y}_t) = (x''_{t-1}, y''_{t-1})$ von y_0 aus startet, mit $t = 2, 3, \dots$

$$w(y') < v\{(x''_t, y''_t)\} + \varepsilon.$$

Daraus folgt mit (3.3)

$$(3.4) \quad \begin{aligned} \varepsilon &= u(x', y') + \delta w(y') - w(y_0) < u(x', y') + \delta(v\{(x''_t, y''_t)\} + \varepsilon) - w(y_0) \\ &= v\{(\tilde{x}_t, \tilde{y}_t)\} + \delta\varepsilon - w(y_0) \leq \delta\varepsilon \end{aligned}$$

im Widerspruch zu $\delta\varepsilon < 0$.

Also gilt für $w(y_0)$ die Funktionalgleichung (vgl. (3.3))

$$(3.5) \quad w(y_0) = \max \{ u(x, y) + \delta w(y) : (x, y) \in T, x \leq y_0 \}.$$

Nach (3.5) existiert zu jedem $\bar{y}_{t-1} \geq 0$ ein $(\bar{x}_t, \bar{y}_t) \in T$

mit $\bar{x}_t \leq \bar{y}_{t-1}$, so daß gilt

$$w(\bar{y}_{t-1}) = u(\bar{x}_t, \bar{y}_t) + \delta w(\bar{y}_t)$$

Daraus erhält man rekursiv für alle $N \in \mathbb{N}$

$$(3.6) \quad w(y_0) = \sum_{t=1}^N \delta^{t-1} u(\bar{x}_t, \bar{y}_t) + \delta^N w(\bar{y}_N).$$

Da w auf der kompakten Menge $Y := \{y : \exists x \text{ mit } (x, y) \in T\}$ stetig ist, folgt durch Grenzübergang $w(y_0) = v(\{\bar{x}_t, \bar{y}_t\})$.

Ist u streng konkav, so ist auch $u + \delta w$ streng konkav und $\{(\bar{x}_t, \bar{y}_t)\}$ ist eindeutig für jedes t .

b) Die Mengen⁴

$$(3.7) \quad Y := \{y : \exists x \text{ mit } (x, y) \in T\}$$

$$(3.8) \quad L(y_0) := \{\bar{y} : \exists \bar{x} \text{ mit } (\bar{x}, \bar{y}) \in T, w(y_0) = u(\bar{x}, \bar{y}) + \delta w(\bar{y})\} \subset Y$$

sind jeweils nichtleer, konvex und kompakt.

Sei $y_0^n \in Y$ mit $y_0^n \rightarrow y_0$ und $y_1^n \in L(y_0^n)$ mit $y_1^n \rightarrow y_1$.

Dann existiert nach (3.5) ein $(x_1^n, y_1^n) \in T$, so daß gilt:

$$w(y_0^n) = u(x_1^n, y_1^n) + \delta w(y_1^n) \quad \text{für alle } n.$$

Gilt $x_1^n \rightarrow x_1$, so erhält man durch Grenzübergang

$$w(y_0) = u(x_1, y_1) + \delta w(y_1)$$

also hat man $y_1 \in L(y_0)$. Nach [8], Seite 65, Definition 4.4, ist L eine abgeschlossene Abbildung auf Y und $L(y_0)$ ist eine nichtleere, konvexe Teilmenge von Y . Dann besagt der Fixpunktsatz von Kakutani, die Abbildung L besitzt einen Fixpunkt, oder, es existiert ein $\bar{y} \in Y$ mit $\bar{y} \in L(\bar{y})$.

Nach (3.8) ergibt sich daraus

$$(3.9) \quad w(\bar{y}) = u(\bar{x}, \bar{y}) + \delta w(\bar{y}) \quad \text{mit } \bar{x} \leq \bar{y} \quad \text{und } (\bar{x}, \bar{y}) \in T.$$

Das stationäre Programm $\{(\bar{x}, \bar{y})\}$ ist optimal.

Die Optimalität von stationären Programmen hängt eng mit folgender Eigenschaft zusammen:

DEFINITION 8:

Ein Prozeß $(\bar{x}, \bar{y}) \in T$ hat die Stützeigenschaft, wenn ein Vektor $q \geq 0$ existiert mit

$$(3.9) \quad q(\bar{y} - \bar{x}) = 0$$

$$(3.10) \quad u(x, y) - u(\bar{x}, \bar{y}) \leq q(x - \bar{x}) - \delta q(y - \bar{y})$$

für alle $(x, y) \in T$.

SATZ 8:

Wenn $(\bar{x}, \bar{y}) \in T$ die Stützeigenschaft besitzt und gilt $\bar{x} \leq \bar{y}$, dann ist das stationäre Programm $\{(\bar{x}, \bar{y})\}$ optimal.

Beweis:

Sei $\{(x_t, y_t)\}$ ein Programm, welches von \bar{y} aus startet. Dann gilt nach Voraussetzung für alle $t \geq 1$

$$u(x_t, y_t) - u(\bar{x}, \bar{y}) \leq q(x_t - \bar{x}) - \delta q(y_t - \bar{y}).$$

Die Multiplikation mit entsprechenden Diskontfaktoren und Summation über t ergibt

$$\begin{aligned} \sum_{t=1}^N \delta^{t-1} u(x_t, y_t) - \sum_{t=1}^N \delta^{t-1} u(\bar{x}, \bar{y}) &\leq \sum_{t=1}^N \delta^{t-1} q(x_t - \bar{x}) - \sum_{t=1}^N \delta^t q(y_t - \bar{y}) \\ &= \underbrace{\sum_{t=1}^N \delta^{t-1} q(x_t - y_{t-1})}_{\leq 0} + \underbrace{\sum_{t=1}^N \delta^{t-1} q(\bar{y} - \bar{x}) + \delta^N q(\bar{y} - y_N)}_{= 0} \\ &\leq \delta^N q(\bar{y} - y_N) \end{aligned}$$

Für $N \rightarrow \infty$ erhält man

$$v\{(x_t, y_t)\} - v\{(\bar{x}, \bar{y})\} \leq 0,$$

also ist $\{(\bar{x}, \bar{y})\}$ optimal.

Zur Veranschaulichung sei nochmals Beispiel 3 mit dem Diskontfaktor $\delta = 3/4$ herangezogen. Die Frage, nach einem stationären Programm $\{(\bar{x}, \bar{y})\}$ ist äquivalent mit der Existenz einer Lösung $q \geq 0$ des Ungleichungssystems:

$$\begin{aligned} 2x - y - 2\bar{x} + \bar{y} &\leq q(x - \bar{x}) - \frac{3}{4}q(y - \bar{y}) \\ q(\bar{x} - \bar{y}) &= 0 \\ \bar{x} &\leq \bar{y} \quad \text{für alle } (x, y) \in T \end{aligned}$$

Durch Umformung und $q \neq 0$ erhält man wegen $\bar{x} = \bar{y}$

$$2x - qx - y + \frac{3}{4}qy = \bar{x} - \frac{1}{4}q\bar{x}.$$

Für $q = 2$ ist jedes stationäre Programm $\{(\bar{x}, \bar{y})\}$ mit $\bar{x} = \bar{y}$ optimal.

LITERATURVERZEICHNIS

- [1] BROCK, W.A.: On Existence of Weakly Maximal Programmes
in a Multisector Economy,
Review of Economic Studies 37, 1970.
- [2] DEBREU, G.: Theory of Value, New York, 1959.
- [3] GALE, D.: On Optimal Development in a Multit-Sector
Economy,
Review of Economic Studies 34, 1967.
- [4] HILDENBRAND, W.: Mathematische Grundlagen zur nichtlinearen
Aktivitätsanalyse, Unternehmensforschung,
Würzburg, 1966.
- [5] KARLIN, S.: Mathematical Methods and Theory in Games,
Programming and Economics, Addison-Wesley,
Reading, Mass., 1959.
- [6] KOOPMANS, D.: Activity Analysis of Production and Allocation
New York, London, 1951.
- [7] MALINVAUD, E.: Capital Accumulation and Efficient Allocation
of Resources, Econometrica 21, 1953.
- [8] NICAIDO, H.: Convex Structures and Economic Theory,
New York, London, 1968.
- [9] OPITZ, O.: Zum Problem der Aktivitätsanalyse, Zeit-
schrift für die gesamte Staatswissenschaft 127
1971.
- [10] RAMSAY, F.P.: A Mathematical Theory of Savings, Economic
Journal 38, 1928.
- [11] STARRETT, D.A. - M. KURZ: On the Efficiency of Competitive
Programmes in an Infinite - Horizon -
Model, Review of Economic Studies 37,
1970.

- [12] SUTHERLAND, W.R.: On Optimal Development in a Multi-Sectoral Economy : the Discounted Case, Review of Economic Studies 37, 1970.
- [13] v. WEIZSÄCKER, C.C.: Existence of Optimal Programmes of Accumulation for an Infinite Time Horizont, Review of Economic Studies 32, 1965.

Zur Existenz von Produktionsfunktionen

von G. Bol, Karlsruhe

In [3] , [5] und [8] wird eine Unternehmung, die aus m Faktorgütern (inputs) n Produktgüter (outputs) herstellt, beschrieben durch eine Funktion $P : V \rightarrow \mathcal{P} X$ ¹⁾ - wobei V eine nichtleere Teilmenge des R_+^m und X nichtleere Teilmenge des R_+^n ist -, die zu einer Kombination $v \in V$ von Inputquantitäten die produzierbaren Kombinationen von Outputquantitäten zuordnet. In [8] wird daneben eine Funktion $F : X \rightarrow \mathcal{P} V$ betrachtet, die einer Kombination von Outputquantitäten die Menge aller Faktorkombinationen zuordnet, bei der sie produziert werden kann. In derselben Arbeit werden Eigenschaften der Funktionen F und P angeführt und diskutiert, außerdem mehrere Folgerungen abgeleitet. Hier wird nun eine kurze Zusammenfassung dieser Axiome angegeben²⁾, einige weitere Folgerungen abgeleitet und ein Kriterium für die Existenz einer Produktionsfunktion im klassischen Sinn bewiesen. Außerdem wird kurz auf das Problem der Stetigkeit bei mengenwertigen Produktionsfunktionen eingegangen.

1) $\mathcal{P} X \equiv \{ M \mid M \text{ Teilmenge von } X \}$.

2) Dazu verwenden wir hier für das Paar (F, P) im obigen Sinn den Ausdruck "Produktionssystem".

Definition 1:

Ein Paar (F, P) von Abbildungen, $F: X \longrightarrow \mathcal{P} V - \{\emptyset\}$,
 $\emptyset \neq X \subset \mathbb{R}_+^n$, $\emptyset \neq V \subset \mathbb{R}_+^m$, $P: V \longrightarrow \mathcal{P} X - \{\emptyset\}$,

heißt Produktionssystem, wenn die folgenden Eigenschaften erfüllt sind:

1. X und V sind kompakt;
2. a) für alle $x_2 \in X$ und $x_1 \in \mathbb{R}_+^n$ mit $x_1 \leq x_2$ ¹⁾
 sei $x_1 \in X$,
 b) für alle $v_2 \in V$ und $v_1 \in \mathbb{R}_+^m$ mit $v_1 \leq v_2$ ¹⁾
 sei $v_1 \in V$;
3. $\forall x \in X, \forall v \in V: x \in P(v) \iff v \in F(x)$;
4. a) $\forall v \in V: P(v)$ abgeschlossen,
 b) $\forall x \in X: F(x)$ abgeschlossen;
5. a) Sei $v_1 \leq v_2, v_1, v_2 \in V$ und $v_1 \in F(x_0)$ für ein
 $x_0 \in X$, dann sei $v_2 \in F(x_0)$,
 b) Sei $x_1 \leq x_2, x_1, x_2 \in X$ und $x_2 \in P(v_0)$ für ein
 $v_0 \in V$, dann sei $x_1 \in P(v_0)$;
6. $P(o) = \{o\}$.

Eine Interpretation dieser Eigenschaften ist in [5] und [8] gegeben. Eigenschaft 6, die hier hinzugefügt wurde, entspricht Eigenschaft 4 in [6], die dort als Nichtexistenz eines Schlaraffenlandes interpretiert wird. Da $P(v)$ nicht-leer ist, folgt aus 5) $o \in P(v)$ für alle $v \in V$, d.h. man

1) Für $x_1, x_2 \in \mathbb{R}_+^k$ sei

$$x_1 \leq x_2 \iff x_{1,i} \leq x_{2,i} \text{ für alle } i = 1 \dots k$$

$$x_1 < x_2 \iff x_{1,i} < x_{2,i} \text{ für alle } i = 1 \dots k$$

kann bei jeder Kombination von Inputquantitäten nichts produzieren. 6. besagt darüber hinaus, bei keinem Einsatz kann nichts produziert werden. Eigenschaft 3. erlaubt im übrigen jede der Abbildungen P und F durch die andere auszudrücken, es ist nämlich

$$\begin{aligned} F(x) &= \{v \in V \mid x \in P(v)\} & \text{und} \\ P(v) &= \{x \in X \mid v \in F(x)\} \end{aligned}$$

Damit ergibt sich für 5. die äquivalente Formulierung:

$$\begin{aligned} 5'. \quad a) \quad & \forall v_1, v_2 \in V: v_1 \leq v_2 \implies P(v_1) \subset P(v_2), \\ b) \quad & \forall x_1, x_2 \in X: x_1 \leq x_2 \implies F(x_2) \subset F(x_1). \end{aligned}$$

In [6] und [10] wird die Menge aller technisch möglichen Produkt-Faktor-Kombinationen als Technologie der Unternehmung bezeichnet. Sei nun (F,P) ein Produktionssystem, so erhält man die zu (F,P) gehörige Technologie als Menge aller Paare (x,v) , die produzierbar sind, d.h. mit $x \in P(v)$.

Definition 2 und Lemma:

Sei (F,P) ein Produktionssystem, dann sei die zugehörige Technologie T definiert durch $T = \{(x, v) \mid x \in P(v), v \in V\}$ und es gilt:

$$T = \{(x,v) \mid v \in F(x)\} = \bigcup_{v \in V} \{v\} \times P(v) = \bigcup_{x \in X} F(x) \times \{x\}.$$

Satz 1:

Sei (F, P) ein Produktionssystem, so ist die zugehörige Technologie T abgeschlossen. Der Beweis beruht auf folgendem Lemma:

Lemma 1:

Sei $(x_n, v_n)_n$ eine Folge in T mit Grenzwert (x_0, v_0) , so existiert eine Folge $(x'_k, v'_k)_k$ in T mit Grenzwert (x_0, v_0) und 1. $x'_k \leq x_0$ 2. $x'_{k_1} \leq x_{k_2}$ für alle $k_1 \leq k_2$.

Beweisskizze (ein ausführlicher Beweis ist in [1] gegeben):

Zunächst existiert eine Teilfolge (x_{n_k}, v_{n_k}) mit

$$x_{n_k, i} > x_{0, i} \text{ für eine feste Menge von Komponenten } \{i_1, \dots, i_m\} \text{ und} \\ x_{n_k, i} \leq x_{0, i} \text{ für } i \notin \{i_1 \dots i_m\}.$$

Dann sei x'_{n_k} definiert durch

$$x'_{n_k, i} = \begin{cases} x_{n_k, i} & \text{für } i \notin \{i_1 \dots i_m\} \\ x_{0, i} & \text{für } i \in \{i_1 \dots i_m\} \end{cases}.$$

Nach 5. ist $(x'_{n_k}, v_{n_k}) \in T$ und aus dieser Folge wählt

man eine Teilfolge mit Eigenschaft 2..

Entsprechend diesem Lemma genügt es, um den Satz zu beweisen, zu zeigen, daß der Grenzwert (x_0, v_0) einer

Folge (x_k, v_k) in T mit den Eigenschaften 1. und 2. in T liegt. Dies erfolgt in zwei Schritten:

Zunächst folgt aus der Abgeschlossenheit von $F(x_n)$ und der Monotonie von (x_n) , daß $v_0 \in F(x_n)$ ist für alle n . Daraus folgt dann wegen der Abgeschlossenheit von $P(v_0)$: $x_0 \in P(v_0)$ und damit $(x_0, v_0) \in P(v_0)$.

In [7] werden für mengenwertige Funktionen die Begriffe "abgeschlossen" und "nach oben halbstetig" eingeführt und gezeigt, daß dies in einigen Fällen äquivalent ist. Da man auf der Menge der kompakten Teilmengen des \mathbb{R}^n eine Topologie erklären kann (vgl. [2], [6]), und wegen 1. und 4. P bzw. F Abbildungen mit Werten dieser Menge sind, kann man weitergehend die Frage der Stetigkeit von P und F untersuchen und Vergleiche mit den oben erwähnten Begriffen ziehen. Dies ist in [1] durchgeführt und es sollen hier nur noch die wichtigsten Definitionen und Sätze angeführt werden.

Definition 3:

Sei M eine nichtleere Teilmenge des \mathbb{R}^n , $\mathcal{K}(M) = \{A \mid \emptyset \neq A \subset M \text{ kompakt}\}$ und $\|\cdot\|$ eine beliebige Norm des \mathbb{R}^n . Auf $\mathcal{K}(M)$ sei die Funktion $d: \mathcal{K}(M) \times \mathcal{K}(M) \longrightarrow \mathbb{R}$ definiert durch:

$$d(A_1, A_2) = \max \left(\sup_{x \in A_1} \delta(x, A_2), \sup_{x \in A_2} \delta(x, A_1) \right)$$

wobei $\delta(x, A_2) = \inf_{y \in A_2} \|x - y\|$ für alle $x \in \mathbb{R}^n$ und

$A_2 \subset \mathbb{R}^n$ ist.

In [2] ist gezeigt, daß d eine Metrik ist. Bezüglich dieser Metrik gilt:

Lemma 2:

Sei P_n eine Folge in $\mathcal{K}(M)$ mit Grenzwert P_0 , dann gilt:

$$P = \{x \mid \forall n \in \mathbb{N} \exists x_n \in P_n \text{ mit } \lim x_n = x\}.$$

Satz 2:

Sei M kompakt, dann ist:

1. $\mathcal{K}(M)$ bezüglich der angegebenen Metrik vollständig, d.h. eine Folge konvergiert genau dann, wenn sie Cauchy-Folge ist.
2. (Auswahlsatz von Hadwiger und Blaschke)
 $(\mathcal{K}(M), d)$ kompakt, d.h. aus jeder unendlichen Teilmenge von $\mathcal{K}(M)$ läßt sich eine konvergente Folge von paarweise verschiedenen Elementen auswählen.

Der Beweis der Behauptungen ist in [2] gegeben.

Definition 4 (vgl. [7]):

$f: D \rightarrow \mathcal{K}M \setminus \{\emptyset\}$, $D \subset \mathbb{R}^m$, $M \subset \mathbb{R}^n$ heißt abgeschlossen in $x_0 \in D$, wenn für alle Folgen x_n in D mit Grenzwert $x_0 \in M$ und alle Folgen y_n in \mathbb{R}^n mit $y_n \in f(x_n)$ und Grenzwert y_0 , $y_0 \in f(x_0)$ ist. f heißt abgeschlossen in D , wenn f abgeschlossen in x_0 für alle $x_0 \in D$.

Folgerung:

Wenn f abgeschlossen in x_0 ist, dann ist offensichtlich $f(x_0)$ abgeschlossen. Demnach ist f eine Abbildung in $\mathcal{K}(M)$, wenn f abgeschlossen in D . Aus Satz 1 folgt unmittelbar, daß für ein Produktionssystem (F, P) F und P abgeschlossene Abbildungen sind.

Definition 5:

Eine Abbildung $f: M \longrightarrow \mathcal{K}(M) - M \quad \mathbb{R}^m, N \subset \mathbb{R}^n$ nichtleer -
heißt nach oben (unten) halbseitig in x_0 , wenn

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \quad \forall x \in M: \|x - x_0\| < \delta \\
\implies \sup_{y \in f(x)} \delta(y, f(x_0)) < \varepsilon \quad \left(\sup_{y \in f(x_0)} \delta(y, f(x)) < \varepsilon \right).$$

f heißt nach oben (unten) halbstetig in M , wenn f nach oben (unten) halbstetig in allen Punkten von M ist.

Diese Definition von nach oben halbstetig fällt mit der Definition in [7] zusammen auf Grund der folgenden Überlegung:

Sei $f(x_0) + \varepsilon = \{x \mid \delta(x, f(x_0)) < \varepsilon\}$, dann ist

$$(*) \quad \sup_{y \in f(x)} \delta(y, f(x_0)) < \varepsilon \iff f(x) \subset f(x_0) + \varepsilon.$$

Eine Umgebung von $f(x_0)$ im Sinne von [7]¹⁾ ist eine Menge U mit: $f(x_0) + \varepsilon \subset U$.

1) Es handelt sich hier nicht um Umgebungen von $f(x_0)$ bzgl. der von d induzierten Topologie.

Aus (*) folgt außerdem: Sei f nach oben halbstetig in x_0 , dann ist für eine Folge x_n mit Grenzwert x_0 , falls $f(x_n)$ konvergiert, $\lim f(x_n) \leq f(x_0) + \varepsilon$ für alle $\varepsilon > 0$ und damit $\lim f(x_n) \leq f(x_0)$.

Satz 3 (vgl. [7]):

Sei $f: M \rightarrow \mathcal{K}(N)$ nach oben halbstetig in x_0 , dann ist f abgeschlossen in x_0 . Wenn N kompakt ist, gilt auch die Umkehrung.

Daraus folgt:

Satz 4:

Sei (F, P) ein Produktionssystem, so sind F und P nach oben halbstetig in X bzw. V .

Stetigkeit von F bzw. P kann aus den Axiomen 1. - 6. nicht gefolgert werden; es ist leicht ein Beispiel anzugeben, indem P und F unstetig sind, aber 1. - 6. erfüllt sind (vgl. [1]).

Eine notwendige und hinreichende Bedingung für Stetigkeit erhält man auf Grund des folgenden Satzes:

Satz 5:

Für ein Produktionssystem (F, P) und $v_0 \in V$ gilt:

1. Für alle Folgen $(v_n)_n$ in V $\lim v_n = v_0$
und $v_{n,i} < v_{0,i}$ für $i = 1 \dots m$ konvergiert $P(v_n)$
2. Der Grenzwert ist für alle Folgen wie in 1. gleich
und ist die abgeschlossene Hülle von $\bigcup_{v < v_0} P(v)$. (Diese

Menge sei mit $\tilde{P}(v_0)$ bezeichnet)

Beweis siehe [1] .

Daraus ergibt sich:

Satz 6:

Sei v_0 innerer Punkt von V , so ist P genau dann stetig in v_0 , wenn $\tilde{P}(v_0) = P(v_0)$.

Ein entsprechendes Kriterium erhält man auch für die Funktion F .

Um die Frage der Existenz einer Produktionsfunktion zu einem Produktionssystem zu untersuchen, werden zunächst einige Begriffe eingeführt:

Definition 6:

$(x, v) \in T$ heißt effizient (technisch optimal)¹⁾, wenn für alle $(x', v') \in T$ mit $x' \geq x$ und $v \leq v'$ folgt $x = x'$ und $v = v'$.

Bemerkung:

Effiziente Produkt-Faktor-Kombinationen aus T sind die maximalen Elemente von T bei der folgenden Relation R :

$$(x, v) R (x', v') \iff x \leq x' \text{ und } v \geq v'.$$

(x', v') wird dem Paar (x, v) bei R vorgezogen, wenn die

1) vgl. [5]

Quantitäten bei allen Produktgütern bei x' größer oder gleich wie bei x bei kleineren oder gleichen Inputquantitäten v' im Vergleich zu v sind.

Definition 7:

Eine Funktion $f : T \rightarrow R$ heißt Produktionsfunktion auf T , wenn $f(x,v) = 0 \iff (x,v)$ effizient.

Wie schon in [6] bemerkt, gilt:

Lemma 3:

Es existiert genau dann eine stetige Produktionsfunktion auf T , wenn die Menge aller effizienten Produkt-Faktor-kombinationen abgeschlossen ist.

Wie man an Beispielen leicht sehen kann, gilt nicht allgemein, daß auf der zu einem Produktionssystem gehörenden Technologie eine stetige Produktionsfunktion existiert. In [5] wird mit Hilfe des Satzes von Debreu über die Darstellbarkeit von Präferenzordnungen durch reellwertige Funktionen ein Kriterium für die Existenz einer stetigen Produktionsfunktion angegeben. Im folgenden soll nun eine hinreichende Bedingung bewiesen werden, die dieses Hilfsmittel vermeidet.

Satz 7:

Sei (F,P) ein Produktionssystem mit den Eigenschaften:

a) Sei $(x_1, \dots, x_n) \in P(v)$ mit:

$\exists k \in \{1 \dots n\} \nexists \lambda \in \mathbb{R}$ mit $\lambda > x_k$ und
 $(x_1, \dots, x_{k-1}, \lambda, x_{k+1}, \dots, x_n) \in P(v)$, dann ist
 (x, v) effizient;

b) Sei $(v_1, \dots, v_m) \in F(x)$ mit
 $\exists k \in \{1 \dots m\} \nexists \lambda \in \mathbb{R}$ mit $\lambda < v_k$ und
 $(v_1, \dots, v_{k-1}, \lambda, v_{k+1}, \dots, v_m) \in F(x)$, dann ist
 (x, v) effizient.

Dann existiert eine stetige Produktionsfunktion.

Der Beweis beruht im wesentlichen auf

Lemma 4:

Sei (F, P) ein Produktionssystem mit den Eigenschaften

a) und b), dann gilt:

$(x, v) \in T$ ist effizient $\iff M_{(x, v)} \cap T = \emptyset$, wobei
 $M_{(x, v)} = \{(\bar{x}, \bar{v}) \mid x \in \mathbb{R}^n, v \in \mathbb{R}^m, \bar{x} > x, \bar{v} < v\}$ ist.

Beweis:

Sei (x, v) effizient, dann existiert kein $(\bar{x}, \bar{v}) \in T$ mit
 $(\bar{x}, \bar{v}) \neq (x, v)$ und $\bar{x} > x, \bar{v} < v$; also ist $M_{(x, v)} \cap T = \emptyset$.

Sei $(x, v) \in T$ nicht effizient. Dann gibt es wegen Eigen-
 schaft a) ein $x' \in X$ mit $(x', v) \in T$ und $x'_1 > x_1, x'_i = x_i$
 für $i = 2, \dots, n$. Sei $x^{(1)}$ definiert durch:

$$x_1^{(1)} = x_1 + \frac{x'_1 - x_1}{2}, \quad x_i^{(1)} = x_i \text{ für } i = 2, \dots, n, \text{ dann}$$

ist $(x^{(1)}, v) \in T$ und nicht effizient. Sei $(x^{(k-1)}, v) \in T$
 nicht effizient mit $x_i^{(k-1)} > x_i$ für $i = 1, \dots, k-1$ und

$x^{(k-1)} = x_i$ für $i = k, \dots, n$, so existiert ein $(\tilde{x}, v) \in T$ mit $\tilde{x}_i = x_i^{(k-1)}$ für $i \neq k$ und $\tilde{x}_k > x_k^{(k-1)}$. Sei nun $x^{(k)}$ definiert durch $x_i^{(k)} = x_i^{(k-1)}$ für $i \neq k$ und $x_k^{(k)} = x_k^{(k-1)} + \frac{x_k - x_k^{(k-1)}}{2}$, so ist $(x^{(k)}, v) \in T$ nicht effizient und

$x_i^{(k)} > x_i$ für $i = 1 \dots k$ und $x_i^{(k)} = x_i$ für $i = k+1, \dots, n$.

Nach n Schritten erhält man ein $x^{(n)}$ mit $(x^{(n)}, v) \in T$ nicht effizient und $x^{(n)} > x$. Entsprechend erhält man nach m Schritten ein $v^{(m)}$ mit $(x^{(n)}, v^{(m)}) \in T$ nicht effizient und $v^{(m)} < v$. Dann ist $(x^{(n)}, v^{(m)}) \in M_{(x,v)}$ und damit $M_{(x,v)} \cap T \neq \emptyset$.

Bemerkung:

Aus diesem Lemma folgt: Hat ein Produktionssystem (F, P) die Eigenschaften a) und b), so gilt:

Sei $x \in P(v)$ maximal bzgl. der Relation " \succ ", so ist (x, v) effizient und damit v minimal in $F(x)$ bzgl. der Relation " \succ ", und umgekehrt (vgl. [8], (2.10)).

Beweis von Satz 7:

Sei $(x^{(n)}, v^{(n)})_n$ eine Folge effizienter Punkte in T mit Grenzwert $(x^{(0)}, v^{(0)})$ und sei $(x^{(0)}, v^{(0)})$ nicht effizient. Dann ist $M_{(x^{(0)}, v^{(0)})} \cap T \neq \emptyset$ und es gibt ein

$(x', v') \in T$ mit $x' > x^{(0)}$ und $v' < v^{(0)}$. Daher ist

$$\varepsilon = \frac{1}{2} \min \{ x'_1 - x_1^{(0)}, \dots, x'_n - x_n^{(0)}, v_1^{(0)} - v'_1, \dots, v_m^{(0)} - v'_m \} > 0$$

und es existiert ein n_0 mit $\|(x^{(n)}, v^{(n)}) - (x^{(0)}, v^{(0)})\| < \varepsilon$ für alle $n \geq n_0$. Dann gilt aber $x^{(n)} \leq x'$ und $v^{(n)} \leq v'$ mit $(x^{(n)}, v^{(n)}) \neq (x', v')$. Dies steht im Widerspruch zur Effizienz von $(x^{(n)}, v^{(n)})$. Also ist $(x^{(0)}, v^{(0)})$ effizient.

Bemerkung:

Aus den Eigenschaften a) und b) folgt auch die Existenz einer i -ten Produktfunktion, bzw. einer i -ten Faktorfunktion für alle i . Sei

$$\begin{aligned} (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n, v) &\in R^{n+m-1} \text{ mit} \\ (x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n, v) &\in T, \text{ so sei} \\ f^i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n, v) &= \\ &\max \{ x_i \in R \mid (x_1, \dots, x_i, \dots, x_n) \in P(v) \}. \end{aligned}$$

Nach Bedingung a) ist dann

$$\begin{aligned} (x_1, \dots, x_{i-1}, \lambda, x_{i+1}, \dots, x_n, v) &\in T \text{ effizient genau dann, wenn} \\ \lambda &= f^i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n, v) \text{ ist.} \end{aligned}$$

Die Bedingungen a) und b) können bei weiterer Gültigkeit von Satz 7 abgeschwächt werden zu

a') Sei $(x_1, \dots, x_n) \in P(v)$
mit $x_i = \max \{ \lambda \mid (x_1, \dots, x_{i-1}, \lambda, x_{i+1}, \dots, x_n) \in P(v) \}$
für ein $i \in \{1, \dots, n\}$, so ist

$$\begin{aligned} 1. \quad x_i &= \max \{ \lambda \mid \exists (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n) \in R^{n-1} \text{ mit} \\ &\quad (y_1, \dots, y_{i-1}, \lambda, y_{i+1}, \dots, y_n) \in X \} \end{aligned}$$

oder

2. (x, v) effizient.

b') Sei $(v_1, \dots, v_m) \in F(x)$ mit

$$\begin{aligned} v_i &= \min \{ \lambda \mid (v_1, \dots, v_{i-1}, \lambda, v_{i+1}, \dots, v_m) \in F(x) \} \\ &\text{für ein } i \in \{1, \dots, m\}, \text{ so ist} \end{aligned}$$

$$1. \quad v_i = 0$$

oder

2. (x, v) effizient.

Der Beweis verläuft analog, verliert aber durch technische Details an Übersichtlichkeit und wurde hier deshalb für die stärkeren Forderungen a) und b) durchgeführt. Hieraus folgt nun für den Fall eines Produktes und eines Faktors das folgende

Korrolar:¹⁾

Sei T konvex, dann existiert eine stetige Produktionsfunktion.

Beweis:

Es genügt zu zeigen, daß für konvexes T a' und b' gelten. Sei $(x, v) \in T$ mit $x = \max \{ y \in R \mid (y, v) \in T \}$, $x \neq \max X$ und (x, v) nicht effizient. Dann existiert $(x', v') \in T$, $(x', v') \neq (x, v)$ mit $v' \leq v$ und $x' > x$. Sei $x'' = \max X$, dann ist mit $v'' \in F(x'')$ $(x'', v'') \in T$ und $x'' > x$. Da $x = \max \{ y \in R \mid (y, v) \in T \}$ ist, muß $x'' > x$ und $x' = x$ sein. Also ist $v' < v$. Da T konvex ist, ist für $\lambda \in [0, 1]$

$$(1 - \lambda)(v', x) + \lambda(v'', x'') \in T.$$

Wegen $v' < v < v''$ existiert ein $0 < \lambda < 1$ mit $(1 - \lambda)v' + \lambda v'' = v$ und $(1 - \lambda)x + \lambda x'' = x + \lambda(x'' - x) > x$ im Widerspruch zur Voraussetzung $x = \max \{ y \mid (y, v) \in T \}$. Also gilt a' . Sei analog $(x, v) \in T$ mit $v = \min \{ y \in R \mid (x, y) \in T \}$, $v \neq 0$ und (x, v) nicht effizient. Dann existiert $(x', v') \in T$ mit $x' > x$, $v' \leq v$ und $(x', v') \neq (x, v)$.

¹⁾ vgl. [6]

Da $v := \min \{ y \in \mathbb{R} \mid (x, y) \in T \}$ und $(x, v') \in T$ ist, gilt $v' = v$ und damit $x' > x$. Da $(0, 0) \in T$ und T konvex ist, ist wegen $\frac{x}{x'} < 1$ $\frac{x}{x'} (x', v) = (x, \frac{x}{x'} \cdot v) \in T$. Dies ist ein Widerspruch zur Voraussetzung $v = \min \{ y \in \mathbb{R} \mid (x, y) \in T \}$. Also gilt b' .

Literatur

- [1] Bol, Georg: Stetigkeit bei mengenwertigen Produktionsfunktionen. Erscheint in: R. Henn: Operations Research-Verfahren, Verlag Anton Hain
- [2] Bol, Gerrit: Über Auswahlssätze
Math. Physikalische Semesterberichte XII (1965)
- [3] Eichhorn, W.: Theorie der homogenen Produktionsfunktionen. Lecture Notes in Operations Research 22, Springer 1970
- [4] Förstner, K. - Henn, R.: Dynamische Produktionstheorie und Lineare Programmierung.
Meisenheim/Glan 1957
- [5] Henn, R. - Opitz O.: Konsum- und Produktionstheorie II, erscheint in Lecture Notes in Operations Research, Springer Verlag.
- [6] Hildenbrandt, W.: Mathematische Grundlagen zur nichtlinearen Aktivitätsanalyse.
Unternehmensforschung 1966
- [7] Nikaido: Convex structures and economic theory. Academic Press 1968
- [8] Opitz, O.: Zum technischen Optimierungsproblem des Unternehmens. Schweizerische Zeitschrift für Volkswirtschaft und Statistik 1970.
- [9] Opitz, O.: Zum Problem der Aktivitätsanalyse.
Zeitschrift für die gesamte Staatswissenschaft, 1971.
- [10] Wittmann, W.: Produktionstheorie,
Springer-Verlag 1968

Erweiterungen des Open Expanding Economy Model von O. Moeschlin, Karlsruhe

Zusammenfassung

In einem um zwei Axiome erweiterten Morgenstern/Thompson Aussenhandelsmodell wird eine Aussage über mögliche Preise für den Fall gemacht, daß weder exportiert noch importiert wird.

Zum Beweis wird ein bekannter Satz aus der Theorie der Linearprogrammierung herangezogen. Abschließend wird übersichtlich auf die Diskussion einer Importbeschränkung eingegangen.

Das von J. v. Neumann in [11] gegebene Modell einer expandierenden Volkswirtschaft hat verschiedene Ergänzungen, Veränderungen und Verallgemeinerungen erfahren, um nur einige zu nennen, z.B. in [4], [9] und [10]. Auch das Open Expanding Economy Model von Morgenstern/Thompson [9] ist zu diesen Verallgemeinerungen zu zählen. Verglichen mit anderen Wachstumsmodelle z.B. [4], [11] sind im Open Expanding Economy Model Importe und Exporte zugelassen, aber auch Gewinne und Verluste. Was die Untersuchungsmethode anbetrifft, so handelt es sich um Anwendung von Sätzen aus der Theorie der Linearprogrammierung.

Um die vorliegende Arbeit selbsttragend zu halten, wird eine kurze Darstellung des Open Expanding Economy Model vorangestellt. Die dabei verwendeten Symbole sollen auch später wiederholt verwendet werden.

Wie in [4] oder [11] sind A , B zwei nichtnegative $m \times n$ Matrizen (Inputmatrix, Outputmatrix); α , sind zwei nichtnegative Zahlen (Wachstums-, Zinsfaktor).

Damit bildet man:

$$\begin{aligned} M_{\alpha} &= B - \alpha A \\ M_{\alpha} &= B - \alpha A \end{aligned}$$

Zusätzlich werden die folgenden Größen eingeführt:

<u>Dimension</u>	<u>Name</u>	<u>Interpretation</u>
1 x n	w^+	Exportvektor
1 x n	w^-	Importvektor
n x 1	p^+	Vektor der Exportpreise
n x 1	p^-	Vektor der Importpreise
n x 1	y	Preisvektor
m x 1	z^+	Gewinnvektor rentierender Prozesse
m x 1	z^-	Verlustvektor nicht rentierender Prozesse
1 x m	t^+	Vektor der oberen Intensitätsgrenze
1 x m	t^-	Vektor der unteren Intensitätsgrenze
1 x m	x	Intensitätsvektor

$I = \{1, \dots, m\}$, $J = \{1, \dots, n\}$ sind Indexmengen zur Bezeichnung der Prozesse (Zeilen) bzw. Güter (Spalten).

Mit Hilfe von 9 Axiomen kann eine Volkswirtschaft wie folgt definiert werden:

- (A1) $xM_{\alpha} = w^+ - w^-$
(Produktion + Import = Verbrauch + Export)
- (A2) $M_{\alpha}y = z^+ - z^-$
(Wert des Outputs + Verlust = Wert des Inputs + Gewinn)
- (A3) $w^+p^+ = w^-p^-$
(Wert des Exports = Wert des Imports)
- (A4) $t^+z^+ = t^-z^-$
(Gewinn - Verlust Gleichgewicht)

- (A5) $xBy > 0$
etwas an Wert wird produziert
- (A6) $t^- \leq x \leq t^+$
Exportvektor liegt zwischen oberer und unterer Intensitätsgrenze
- (A7) $p^+ \leq y \leq p^-$
Inlandpreis liegt zwischen Import und Exportpreis
- (A8) $z^+ \cdot z^- = 0$
kein Prozess macht sowohl Gewinn als auch Verlust
- (A9) $w^+ \cdot w^- = 0$
kein Gut wird sowohl importiert als auch exportiert

wobei $w^+, w^-, z^+, z^- \geq 0$.

Für eine Diskussion der Axiome (A1) - (A7) halte man sich an [9] S. 448, dort wird auch auf die Beziehungen zu den Modellbedingungen in [4] eingegangen.

Um den Nachweis der Existenz einer Lösung zu (A1) - (A7) mit $\alpha = \beta$ (in [9] fehlen die Axiome (A8), (A9) wird eine Schar zueinander dualer Linearprogramme betrachtet:

$$\begin{array}{llll}
 \text{Min} & -w^+p^+ + w^-p^- & & \\
 xM_\alpha & -w^+ + w^- & = & 0 \\
 -x & & \geq & -t^+ \\
 x & & \leq & t^- \\
 w^+, w^- & & \geq & 0
 \end{array} \quad (1)$$

$$\begin{array}{llll}
 \text{Max} & -t^+z^+ + t^-z^- & & \\
 M_\alpha y & -z^+ + z^- & = & 0 \\
 -y & & \leq & -p^+ \\
 y & & \geq & p^- \\
 z^+, z^- & & \geq & 0
 \end{array} \quad (2)$$

Unter den Voraussetzungen B1 - B4

$$\begin{aligned} \text{B1 :} & \quad 0 \leq p^+ \leq p^- \\ \text{B2 :} & \quad 0 \leq t^- \leq t^+ \\ \text{B3 :} & \quad t^- B p^+ > 0 \\ \text{B4 :} & \quad t^- A p^- > 0 \end{aligned}$$

kann gezeigt werden, daß mindestens ein α existiert, so daß der gemeinsame optimale Funktionswert der Zielfunktion aus (1) und (2) gleich null ist, womit die Existenz einer Lösung des Modells nachgewiesen ist.

Die zusätzlichen Axiome A8, A9

Wie schon erwähnt werden (A8), (A9) in [9] nicht als Axiome aufgeführt, vielmehr werden sie dort als Folgerungen aus (A1) - (A7) betrachtet. Was die Widerspruchsfreiheit des erweiterten Axiomensystems anbetrifft, so existiert unter den Lösungen zu (A1) - (A7) immer eine solche die auch (A8), (A9) erfüllt. Andererseits ist es leicht eine Lösung zu finden, die (A1) - (A7) erfüllt, (A8), (A9) aber nicht:

Nach B1 darf angenommen werden, daß

$$h \in J: \quad p^{+h} = p^{-h}$$

Liegt nun eine Lösung zu (A1) - (A7) mit w^{+h} , w^{-h} vor, so bildet man eine neue Lösung einzig dadurch, daß man $c > 0$ zu w^{+h} , w^{-h} dazuaddiert, d.h. man bildet: w^{+h+c} , w^{-h+c} ; während die übrigen Größen dieselben bleiben. Nach Konstruktion erfüllt die neue Lösung die Axiome (A1) - (A7), während (A9) nicht mehr erfüllt ist. Dies bedeutet, daß auch das erweiterte Axiomensystem unabhängig ist. Für eine ausführliche Durchrechnung halte man sich an [7].

Eine Erweiterung eines Satzes von Morgenstern-Thompson

Sei (\bar{x}, \bar{y}) eine Lösung zu (A1) - (A9) mit $\bar{\alpha} = \bar{\beta}$, so daß
 $(\bar{w}^{+j}, \bar{w}^{-j}) = (0, 0) \quad j \in K \subset J$. Dabei gelte: $p^{+j} < p^{-j}$, $j \in K$;

zudem sei $t^{-i} < t^{+i}$, $i \in L$, $\emptyset \neq L \subset I$.

Der folgende Satz nennt Bedingungen dafür, daß (\bar{x}, \bar{y}) ebenfalls eine Lösung zu (A1) - (A9) mit $\bar{\alpha} = \bar{\beta}$ ist. Dabei werden weitere Lösungen herangezogen: (\bar{x}, \hat{y}) , (\bar{x}, \tilde{y}) wobei \hat{z} , \tilde{z} die zugehörigen Gewinne, Verluste kennzeichnen:

Satz:

(\bar{x}, \bar{y}) ist eine Lösung zu (A1) - (A9) mit $\bar{\alpha} = \bar{\beta}$, wenn y gemäß K1, K2 bestimmt werden.

K1: $y^j = p^{+j}, p^{-j}$ bzw.,
 wenn $\bar{w}^{+j} > 0$, $\bar{w}^{-j} > 0$ bzw. $j \in J - K$

K2: $y^j \in [p^{+j}, p^{-j}]$, $j \in K$, so daß für $\forall i \in L$:
 $z^{+i} > 0$, $z^{-i} = 0$, wenn eine Lösung
 (\bar{x}, \hat{y}) existiert mit $(\hat{z}^{+i} > 0, \hat{z}^{-i} = 0)$

oder

$z^{+i} = 0$, $z^{-i} > 0$, wenn eine Lösung
 (\bar{x}, \tilde{y}) existiert mit $(\tilde{z}^{+i} = 0, \tilde{z}^{-i} > 0)$

oder

$(z^{+i}, z^{-i}) = (0, 0)$.

Beweis:

Zum Beweis (vergl. [5]) muss lediglich geprüft werden, ob die gemäß K1, K2 gefundenen Werte optimale Lösungen der Linearprogramme (1), (2) mit dem fixierten Parameter $\alpha = \bar{\alpha}$ sind. Dies kann mit Hilfe des Komplementaritätssatzes

geschehen; wonach die einzelnen Summanden aus (3) und (4) genau dann null sind, wenn optimale Lösungen der Programme (1), (2) eingesetzt werden:

$$\begin{array}{ll} y(xM_{\bar{\alpha}} - w^+ + w^-) & \text{a.} \\ +z^+(-x + t^+) & \text{b.} \\ +z^- (x - t^-) & \text{c.} \end{array} \quad (3)$$

bzw.

$$\begin{array}{ll} x(M_{\bar{\alpha}}y - z^+ + z^-) & \text{a.} \\ +w^+(-y + p^+) & \text{b.} \\ +w^- (y - p^-) & \text{c.} \end{array} \quad (4)$$

Zu (3): Der Summand a. ist null. Die Summanden b. und c. sind es ebenfalls, denn wird z.B. eine Lösung $z^{+i} > 0$, $z^{-i} = 0$, $i \in L$ bestimmt, dann existiert nach K2 eine Lösung $z^{+i} > 0$, $z^{-i} = 0$, $i \in L$; d.h. aber daß $(-\bar{x} + t^{-i}) = 0$ ist. Daher ist $z^{+i}(-\bar{x}^{+i} + t^{+i}) = z^{-i}(x^{-i} - t^{-i}) = 0$.

Zu (4): Der Summand a. ist null. Die Summanden b. und c. sind es ebenfalls; für $j \in J - K$ folgt dies aus K1; für $j \in K$ ergibt sich die aus der Beschreibung der Lösung (\bar{x}, \bar{y}) . (Insbesondere wird klar, daß die y^j nicht notwendig am Rande des Intervalls $[p^{+j}, p^{-j}]$, $j \in K$ zu liegen brauchen)

Bemerkung:

Für $K = \emptyset$ ergibt sich die in Theorem 4 in [9] gegebene Charakterisierung von Modell-Lösungen.

Ökonomisch besagt der Satz, daß der Preis einer Gutes als Import- bzw. Exportpreis festgelegt ist, je nach ob es sich um ein Import- bzw. Exportgut handelt. Wird ein Gut weder

exportiert noch importiert, so ist sein Preis in einem gewissen Rahmen frei, muß aber so angelegt sein, daß eine bestehende "Gewinn-Verlust-Struktur" nicht gestört wird; d.h. durch die Veränderung eines nicht festgelegten Preises darf ein Prozess der vor der Veränderung mit Gewinn betrieben wurde, nach der Veränderung nicht mit Verlust arbeiten und umgekehrt.

Diskussion einer Importbeschränkung

Im Modell von Morgenstern/Thompson bestehen keine direkten Restriktionen für w^+ , w^- , z^+ , z^- ; d.h. jede Nachfrage kann nötigenfalls durch Import gedeckt, jeder Überschuss kann exportiert werden. Da w^+ , w^- , z^+ , z^- keinen weiteren Beschränkungen unterworfen werden, weisen die Programme (1), (2) für alle $\alpha > 0$ zulässige Lösungen auf. Daher hängt der gemeinsame Funktionswert der Zielfunktionen in (1) und (2) stetig von α ab. Die Stetigkeit wird beim Existenznachweis in [9] wesentlich mitbenutzt.

In [6] werden zusätzliche Restriktionen eingeführt, die als Importbeschränkungen gedeutet werden. Von speziellen Situationsannahmen ausgehend werden dort zusätzliche Bedingungen genannt und diskutiert, die die Existenz von Lösungen im veränderten Modell sichern.

Literaturverzeichnis

- [1] K. Förstner, "Wirtschaftliches Wachstums bei vollständiger Konkurrenz und linearer Technologie", in Operations Research Verfahren : Herausgeber: R. Henn, Bd. 4, (1967).
- [2] R. Henn, "Makroökonomische Expansionsmodelle: v. Neumannsche Interpretation einiger Wachstumsmodelle. Unternehmungsforschung", Bd. 6, Heft 1, 16-25, (1962).
- [3] J. Hülsmann, V. Steinmetz: "A note on the nonexistence of optimal price vectors in the general balanced-growth model of Gale", erscheint in Econometrica.
- [4] Kemeny, J.G., O. Morgenstern und G.L. Thompson, "A Generalization of the von Neumann Model of an Expanding Economy", Econometrica, 24, 115-135, (1956).
- [5] O. Moeschlin, "Zur Eindeutigkeit von Lösungen des Morgenstern/Thompson Aussenhandelsmodells", erscheint in Zeitschrift für die gesamten Staatswissenschaften.
- [6] O. Moeschlin, "Eine Importbeschränkung im Aussenhandelsmodell von Morgenstern/Thompson", erscheint in Zeitschrift für Nationalökonomie.
- [7] O. Moeschlin, B. Rauhut, "Eine Bemerkung zum Axiomensystem des Morgenstern/Thompson Aussenhandelsmodells", erscheint in Zeitschrift für Operations Research (Unternehmungsforschung).

- [8] Morgenstern O. und G.L. Thompson
"Private and Public Consumption and Savings in the
von Neumann Model of an Expanding Economy",
Kyklos, 20, 387-409 (1967).
- [9] Morgenstern O. und G.L. Thompson
"An Open Expanding Economy Model", Naval Research
Logistics Quaterly, Vol. 16, No. 4, 443-457 (1969).
- [10] Morishima, M.
"Economic Expansion and the Interest Rate in
Generalized v. Neumann Models", Econometrica 28,
352-363, (1960).
- [11] v. Neumann, J.
"Über ein ökonomisches Gleichungssystem und eine
Verallgemeinerung des Brouwerschen Fixpunktsatzes",
Ergebnisse eines mathematischen Kolloquiums, No. 8,
73-83 (1937).

Stopp-Probleme und Markoff-Modelle

Linear Programming Algorithms for the Deterministic Discrete Dynamic Programming Problem

by Y. M. I. Dirickx, Berlin

1. The Model

The model discussed in this paper is known as the deterministic discrete dynamic programming model. This model has a simple and useful representation as an optimization problem in a finite network, cf. (13).

Consider a directed network (S,A) , where the finite node set S corresponds to the state space, and the arc set A is a subset of $S \times S$; an arc (s,t) corresponds to being in state s and selecting a decision that causes transition to state (node) t at the next epoch. The reward r_{st} of this arc is the immediate reward for being in state s and selecting that decision.

Let $A(s)$ be the set of arcs emanating from node s . The assumption that each state has at least one decision transcribes to

Assumption $A(s)$ is non-empty for each s in S .

A finite path, say p_k , is a sequence $\{(s_0, s_1), \dots, (s_{k-1}, s_k)\}$ with $(s_i, s_{i+1}) \in A$ for $i = 0, 1, \dots, k-1$; the subscript k indicates the length of the path, which is defined as the number of arcs in the path. If $s_0 = s_k$ in p_k then we call this path a cycle. Cycles will be denoted by Greek letters such as Γ_k , k refers to the length of the cycle. An infinite path, usually denoted by p , is a sequence $p = \{(s_0, s_1), (s_1, s_2), \dots\}$ with $(s_i, s_{i+1}) \in A$ for $i = 0, 1, \dots$, we set the length of this path equal to $+\infty$.

With each path in the network is associated an income stream corresponding to the arcs in the path; the rewards of these arcs are earned in later epochs and are evaluated in terms of their present values. The present value of a finite path p_k is given by

$$r(p_k) = \sum_{i=0}^k \alpha^i r_{s_i, s_{i+1}},$$

where α is the discount factor reflecting the time preferences of the decision maker. The present value of an infinite path p could then be defined as a limiting value of the present values of the sequence of

finite paths $r(p_k)$ for $k = 0, 1, 2, \dots$, where p_k is the k -step truncation of p . This limiting process can be applied without much difficulties when $0 \leq \alpha < 1$; however, if $\alpha \geq 1$ one can hardly expect $r(p_k)$ to converge to a finite value as k approaches infinity.

In this paper we will present a unified treatment for all values $\alpha \geq 0$. To do so, we introduce the notion of equivalent average return, which is defined for a given finite path p_k as the value

$$(1) \quad \bar{r}(p_k) = r(p_k) (1 + \alpha + \dots + \alpha^{k-1})^{-1}.$$

Equivalent average return is a recurring notion in dynamic programming that has proven useful when studying the limiting behavior of discounted income streams ($\alpha < 1$) as the discount factor approaches one from below (cf. (15)). This notion is useful here for other reasons. First, the ordering that results from using (1) for a given finite horizon k is identical with the present value method. Second, as the horizon length becomes unbounded the equivalent average return does not explode when $\alpha \geq 1$ though it might oscillate. To verify this, let $M = \max_{(s,t) \in A} |r(s,t)|$; then (1) im-

plies that

$|\bar{r}(p_k)| \leq M$ for all k . So that we can define the

equivalent average return of an infinite path as

$$(2) \quad \bar{r}(p) = \limsup_{k \rightarrow \infty} \bar{r}(p_k),$$

where p_k is the k -step truncation of p .

There is no compelling justification for using "lim-sup" rather than "lim-inf" in (2), and the two will actually rarely be equal.

Since a policy specifies all future decisions for each starting state, it is easy to see that every non-randomized policy gives rise to an infinite path from each starting state. To avoid unnecessary notation we will fix the initial state and develop the theory by concentrating on infinite paths originating in the particular node. The modifications needed to specify a policy will be noted whenever necessary.

If we define

$$(3) \quad f(s) = \sup_{(p | s_0=s)} \bar{r}(p),$$

where the supremum is taken over all infinite paths with initial node $s_0=s$, then we can state the following infinite horizon problem

- (4) For a given $\alpha \geq 0$ find a path p^* starting in node s such that $\bar{r}(p^*) = f(s)$.

If $0 \leq \alpha < 1$ then it is clear that $r(p_k)$ converges as $k \rightarrow \infty$, however if p^* is optimal with respect to (4) its present value is also maximal. If $\alpha=1$ (the averaging case), then (2) reduces to the usual definition of the average return.

In the next section we will summarise the results concerning the structure of optimal policies to (4). The linear programming algorithms are discussed in Section 3.

2. The Structure of Optimal Policies.

First we introduce some definitions. A finite or infinite path is stationary if every time the same node reappears the same arc is selected. Hence, if an infinite path is stationary, then it consists of an initial stationary path u_1 , possibly zero length, and a stationary cycle γ_k repeated infinitely often with only one node in common with u_1 , so $p = \{u_1, \gamma_k, \gamma_k, \dots\}$.

A stationary policy can be described as a mapping from the state space S into A . Application of a stationary policy leads to a stationary infinite path for each starting state, the converse is not necessarily true. For the case $0 \leq \alpha < 1$ we have the following well-known result, cf. (2), (6), (13).

Theorem 1 If $0 \leq \alpha \leq 1$, then there exists a stationary policy which is optimal with respect to (5).

If $\alpha > 1$, the situation is more complex; the pathologies resulting from a discount factor greater than one are discussed in (7), the main result derived there is summarized in the following

Theorem 2 Suppose $\alpha > 1$. For every given node s in S there exists a path satisfying (5) of the form

$$p^* = \{u_1, (p_n, \gamma_k^{(n)}, q_\ell)_{n=1,2,\dots}\}.$$

where u_1 is an arbitrary path joining the initial node to the last node of q_ℓ and p_h is an arbitrary stationary path joining the last node of q_ℓ to the cycle γ_k . The cycle γ_k and the path q_ℓ are both stationary and have only one node in common. The notation $\gamma_k^{(n)}$ indicates that the cycle γ_k is repeated n times.

Moreover,

$$(5) \quad \bar{r}(p^*) = \alpha^{-\ell} (\alpha - 1) \{r(\gamma_k) (\alpha^\beta - 1)^{-1} + r(q_\ell)\} \quad (*).$$

(*) A modified definition of equivalent average return was used in (7) and the formulas obtained there are slightly different.

3. Linear Programming Algorithms.

We will first consider the case $0 \leq \alpha < 1$. An algorithm for the averaging case ($\alpha=1$) will be derived from that developed for $\alpha > 1$.

Consider the following linear program for any value $0 \leq \alpha < 1$.

$$\begin{array}{ll}
 \text{Program 1} & \max \sum_{(s,t) \in A} r_{st} x_{st} \\
 & \text{subject to} \\
 & \sum_{\{t; (s,t) \in A\}} x_{st} - \alpha \sum_{\{t; (t,s) \in A\}} x_{ts} = a_s \\
 & \text{for each } s \in S, \\
 & \text{all } x_{st} \geq 0.
 \end{array}$$

The right hand side values a_s are strictly positive for each s , but otherwise arbitrary. Program I was first introduced by D'Epenoux (5) for the Markovian decision problem (the stochastic generalisation of our model). He proved

Theorem 3 Program I computes an optimal stationary policy.

In view of Theorem 1 this implies that Program I

computes an optimal policy, this result is widely known cf. for instance Beckmann (1), Derman (6), Wagner (15).

We can now turn to a discussion of the case $\alpha > 1$; the results reported here are discussed in detail in (8).

Consider any network (S, A) satisfying the assumption made in Section 1, so we can formulate

$$\begin{aligned}
 &\text{Program II} \quad \max \quad \sum_{(s,t) \in A} r_{st} x_{st} \\
 &\quad \text{subject to} \\
 (6) \quad &x_{s0} + \sum_{\{t; (s,t) \in A\}} x_{st} - \alpha \sum_{\{t; (t,s) \in A\}} x_{ts} = 0 \\
 &\text{for each } s \in S, \\
 (7) \quad &\sum_{s \in S} x_{s0} = \alpha - 1 \\
 &\text{all } x_{st} \geq 0, x_{s0} \geq 0.
 \end{aligned}$$

Program II is akin to Program I and to the linear program derived in the dynamic programming literature when $\alpha = 1$. The basic idea is to generate flow in a cycle and to remove one unit of flow from the network. The linear program is a case of the "network with gains" model discussed in Jewell (10), where the α in Program II corresponds to the "gain rate". Since the introduction of a gain rate destroys the usual conservation of flow

(cf. Jewell (10), p. 480), an additional node, called the 0-node, is introduced.

Equations (6) are then the "conservation of flow" equations, and (7) is a normalization constraint needed to keep the amount of flow bounded. Any positive constant works but $\sum_{s \in S} x_{s0} = \alpha - 1$

simplifies the interpretation of optimal solutions to Program II.

The properties of basic solutions to Program II are analysed in the following theorem. Active basic arcs are defined as those associated with strictly positive basic variables.

Theorem 4 The active basic arcs of each basic solution to Program II form one stationary cycle and a unique stationary path leading from that cycle to the 0-node with only one node in common with that cycle.

The proof of this theorem is contained in (8).

If we denote the cycle identified by an optimal basic solution of Program II as γ_k , and the path leading from that cycle to the node t_ℓ by q_ℓ , then we can prove

Corollary 1 The value of the objective function associated with an optimal basic solution to Program II is

$$\alpha^{-k}(\alpha-1) \{ r(\gamma_k) (\alpha^{k-1})^{-1} + r(q_k) \}.$$

In view of (5) in Theorem 2, Corollary 1 seems to suggest that we could apply Program II to the network characterising our dynamic programming problem, there is however a basic difficulty. Application of Program II might yield an optimal solution such that there is no path from the last node of the stationary path to the cycle identified by Program II, so that a policy of the form described in Theorem 2 cannot be found. To circumvent this problem we will identify the ergodic sets of the underlying network and apply Program II to each of these ergodic sets, once this is achieved, the identification of an optimal policy will become trivial. Before describing the algorithm some well-known notions are defined.

A subset \bar{S} of nodes in a network (S,A) is said to be an ergodic set if and only if \bar{S} is connected and if all the nodes of every given path joining any two nodes in \bar{S} are elements of \bar{S} . This definition implies that if a node belongs to an ergodic set, then there is no arc from that node that leads to a node that is not an element of that ergodic set. A node s in S is said to be absorbing if and only if $(s,s) \in A$ and $(s,t) \in A$ for all nodes $t \neq s$. If s is absorbing then it forms an ergodic set of one single element.

Since the algorithm will successively delete sets of arcs and nodes, the assumptions made in Section 1 will not necessarily be met. A node becomes incompatible if positive flow through that node becomes impossible, that is, either there are only arcs leading to or leaving from that node. An incompatible arc arises when one of its end nodes are deleted from the network. Finally, we denote the set of all nodes from which there is a path leading to a node in some set S by the symbol (S) . This set (S) may be computed by augmenting S repeatedly with any node not in S such that an arc $(s,t) \in A$ exists with $t \in S$.

We can now state the algorithm, each of the iterations will be discussed in detail afterwards.

ALGORITHM

- Step 0 Set $(S,A) = (S^0,A^0)$, go to Step 1 with $i = 0$.
- Step 1 Find an ergodic set \bar{S}^1 and the set of arcs \bar{A}^1 , connecting the nodes in \bar{S}^1 , go to Step 2.
- Step 2 Apply Program I to network (\bar{S}^1,\bar{A}^1) . Let $f(\bar{S}^1)$ be its objective function value. Its stationary cycle and stationary path are denoted by γ_{k_1} and α_{k_1} , go to Step 3.
- Step 3 Consider $(S^1-\bar{S}^1,A^1-\bar{A}^1)$, indentify the

incompatible arcs and nodes S_I^i and A_I^i ,
 define $S^{i+1} = S^i - (\bar{S}^i \cup S_I^i)$ and
 $A^{i+1} = A^i - (\bar{A}^i \cup A_I^i)$. If S^{i+1} is empty
 go to Step 4, otherwise go to Step 1 with
 $i + 1 = i$.

Step 4

Rank the ergodic set in decreasing order
 of their objective function,

$f(\bar{S}^{i_1}) \geq \dots \geq f(\bar{S}^{i_m})$. Set $j = 1$, indent-

ify for all $s \in (S^{i_j})$ a path

$\{u_1, (p_{h_j}, \gamma_{k_j}^{(n)}, g_{l_j})_{n=1,2,\dots}\}$, where

u_1 is a path joining node s to the last
 node of q_{l_j} . If $S - \bigcup_{r=1}^j (\bar{S}^{i_r})$ is empty,

STOP; otherwise repeat with $j + 1 = j$.

We now discuss the iteration steps in more detail.
 Any of the algorithms to find ergodic sets might be used
 in Step 1, the Fox-Landi algorithm (9) seems to be most
 suitable for our purposes, cf. also Uebe (14).

The identification of incompatible arcs and nodes
 in Step 3 is not necessary, however it allows for a
 reduction of the linear program to be solved at the next
 iteration.

To obtain all the elements connected to some set \bar{S}^i_j in Step 4, we just analyze sequentially an incidence matrix in which each ergodic set is collapsed to one node. The paths specified in Step 4 are optimal by our previous observations. Note that the above algorithm is computationally efficient; the identification of the ergodic sets is a fast operation, and the linear programs decrease in size when more repetitions are necessary. Moreover Jewell's algorithm for networks with gains (cf. (10) and (12)) can be used to compute optimal basic solutions to Program I.

We can now consider the case $\alpha=1$. If we let $\alpha \rightarrow 1^+$ in Program II, then $x_{s0} \rightarrow 0$ and the following limiting program is obtained

Program III

$$\max \sum_{(s,t) \in A} r_{st} x_{st}$$

subject to

$$\sum_{\{t; (s,t) \in A\}} x_{st} - \sum_{\{t; (t,s) \in A\}} x_{ts} = 0$$

for each $s \in S$

$$(8) \quad \sum_{(s,t) \in A} x_{st} = 1$$

$$\text{all } x_{st} \geq 0$$

The constraint (8) is added to obtain a bounded program. Program III was first derived by Manne (11) and de Ghellinck (3), and computes the best average return stationary cycle as shown by Denardo (4). Hence, in view of Theorem 1, Program III can be used to compute an optimal policy. In general a sequential application of Program III will be needed, the details are left to the reader.

References

- (1) Beckmann, Martin J., Dynamic Programming of Economic Decisions, Springer Verlag, 1968.
- (2) Blackwell, David, "Discrete Dynamic Programming" Annals of Mathematical Statistics, Vol. 33 (1962), pp. 719-726.
- (3) de Ghellinck, G., "Les Problèmes de Décisions Séquentielles, "Cahiers Centre D'Etudes Recherche Opérationnelle, Vol. 2 (1960), pp. 161-179
- (4) Denardo, Eric V., "On Linear Programming in a Markov Decision Problem", Management Science, Vol. 16 (1970), pp. 281-288.
- (5) D'Epenoux, F., "Sun un Problème de production et de stockage dans l' aléatoire", Revue Française de Recherche Operationelle, Vol. 14 (1960) pp. 3-16.

- (6) Derman, C., Finite State Markovian Decision Processes, Academic Press, 1970.
- (7) Dirickx, Yvo M.I., "Deterministic Discrete Dynamic Programming with Discount Factor Greater than One - Structure of Optimal Policies", to appear in Management Science.
- (8) Dirickx, Yvo M.I., "Algorithms for the Deterministic Discrete Dynamic Programming Model with Discount Factor Greater than One", submitted to Operations Research.
- (9) Fox, B. and Landi, "An Algorithm for Identifying the Ergodic Subchains and Transient States of a Stochastic Matrix", Communications of the ACM, Vol. 11 (1968), pp. 619-621.
- (10) Jewell, W.S., "Optimal Flows Through Networks with Gains," Operations Research, Vol. 10 (1962), pp. 476-499.
- (11) Manne, A.S., "On Linear Programming in a Markov Decision Problem", Management Science, Vol. 16 (1970), pp. 281-288.
- (12) Minieka, Edward, "Selected Graph Theoretic Optimization Problems", Chapter 1, Ph.D Thesis, Department of Administrative Sciences, Yale University, 1970.

- (13) Shapiro, Jeremy, "Shortest Route Methods for
Finite State Space Deterministic Dynamic
Programming Problems", SIAM Journal of
Applied Mathematics, Vol. 16.

 - (14) Uebe, Götz, "Classifying the States of a
Finite Markov Chain", these Proceedings.

 - (15) Wagner, Harvey M., "Principles of Operations
Research", Prentice-Hall, 1960.
-

Über Stop-Probleme bei diskreten Markoff-Ketten

von B. H. Goldstein, Karlsruhe

1. Die Arbeit behandelt eine Verallgemeinerung einfacher Stop-Probleme bei diskreten Markoff-Ketten (MK). Als einfaches Stop-Problem soll hier die folgende Aufgabenstellung bezeichnet werden (vergl. Breiman [2], Derman [3], Dynkin-Juschkevitsch [5]): Gegeben sei eine diskrete MK, d.h. eine MK mit diskreter Zeit, höchstens abzählbarem Zustandsraum E und stationären Übergangswahrscheinlichkeiten, ferner eine beschränkte Funktion $a: E \rightarrow \mathbb{R}$. Wird die MK im Zustand $i \in E$ gestoppt, so erhält man als Auszahlung $a(i)$. Gesucht ist eine Stop-Vorschrift, die die zu erwartende Auszahlung maximiert.

Im Hinblick auf Anwendungen ist jedoch die Kenntnis der Übergangsmatrix häufig nicht gegeben. Daher soll hier nur ein Vorwissen in der Form vorausgesetzt werden, daß die Übergangsmatrix, nach der die MK abläuft, Element einer vorgegebenen Menge von stochastischen Matrizen ist. Auf diese Weise kann man Unsicherheiten bzgl. der Kenntnis der Matrizen berücksichtigen. Die sich nun ergebende Problemstellung läßt sich für jeden Startpunkt $i \in E$ als ein Zwei-Personen-Nullsummenspiel (ZPNS) auffassen. Die Arbeit befaßt sich mit diesen Spielen. Es wird eine Übersicht über die wesentlichen Resultate gegeben, Beweise sind in skizzierter Form angegeben (zu Details vergl. [7]).

2. Es sollen hier nur endliche absorbierende MK betrachtet werden, daher sei E eine nicht leere, endliche Menge, $P: E \times E \rightarrow [0,1]$ eine stochastische Matrix, $(x(t), E, P)$ die zugehörige diskrete MK. Man setzt

$$T(P) = \left\{ i : \sum_{t=0}^{\infty} P^t(i, i) < \infty \right\},$$

$$S(P) = \{ i : P(i, i) = 1 \}; \quad E = T(P) \cup S(P),$$

d.h. $T(P)$ ist die Menge der transienten, $S(P)$ die der absorbie-

renden Zustände. Schließlich werde mit P_i die Wahrscheinlichkeitsverteilung bzgl. P und der Einpunktverteilung auf $i \in E$ bezeichnet, mit $E_{i,P} f(x(\tau))$ der Erwartungswert von $f: E \rightarrow R$ bzgl. der Verteilung von $x(\tau)$, τ eine Stop-(Markoff-) Zeit. Dabei soll gelten

$$E_{i,P} f(x(\tau)) = E_{i,P} \left[f(x(\tau)), \tau < \infty \right] \quad \text{sowie}$$

$$E_{i,P} f(x(\infty)) = 0.$$

Man definiert

Definition (2.1)

Es sei X die Menge aller Stopzeiten, $Y \neq \emptyset$ und

$$Y \subseteq \{P: P: E \times E \rightarrow [0,1], P1 \leq 1\}, \quad S(P) = S$$

für alle $P \in Y$, $a: E \rightarrow R$ beschränkt,

$$K_i(\tau, P) = E_{i,P} a(x(\tau)) \quad (\tau \in X, P \in Y). \quad \text{Dann heißt}$$

$\Gamma = (X, Y, K) = (\Gamma(i))$ mit $\Gamma(i) = (X, Y, K_i)$ ($i \in E$) ein Stop-Spiel.

Zu jedem Startpunkt $i \in E$ erhält man also ein ZPNS $\Gamma(i) = (X, Y, K_i)$.

$K_i(\tau, P)$ gibt die Auszahlung an Spieler 1 (den Entscheidenden) von Spieler 2 (Umwelt) an für die Strategiewahl $\tau \in X$ und $P \in Y$. Der daraus resultierende Prozess ist eine zum Zeitpunkt τ gestoppte diskrete MK.

Zunächst werden zwei Definitheitsbedingungen für die Spiele $\Gamma(i)$ hergeleitet.

Satz (2.2)

Es sei $\Gamma = (X, Y, K)$ ein Stop-Spiel mit $a(j) \geq 0$ ($j \in S$).

Dann ist die diskret gemischte Erweiterung

$$\Gamma^d(i) = (X_d, Y_d, K_i) \quad \text{von} \quad \Gamma(i) = (X, Y, K_i) \quad \text{definit,}$$

falls eine der folgenden Bedingungen erfüllt ist:

(1) Zu jedem $\varepsilon > 0$ existiert $n(\varepsilon)$ mit
 $P_i(x(t) \notin S, t > n(\varepsilon)) < \varepsilon \quad (P \in Y)$

(2) Es existiert eine Konstante $C > 0$ mit

$$\sum_{t=0}^{\infty} P^t(j, k) \leq C \text{ für } k \notin S, P \in Y.$$

Beweis:

Es sei $\mathcal{T} = \inf \{t \geq 1 : x(t) \in S\}$ der Zeitpunkt der Absorption für die MK. Wegen $a(j) \geq 0 \ (j \in S)$ kann man X ersetzen durch $\bar{X} = \{\tau \in X : \tau \leq \mathcal{T}\}$, ferner ist offenbar nur der Fall $i \in S$ interessant, es sei daher im folgenden stets $i \in S$. Ist die Bedingung (1) erfüllt, so existiert zu $\varepsilon > 0$ ein $n(\varepsilon)$ mit $P_i(\mathcal{T} > n(\varepsilon)) \leq \varepsilon/2 \sum_E |a(j)| \ (P \in Y)$. Man rechnet leicht

nach, daß dann für $\tau \in \bar{X}$, $\tau' = \min \{\tau, n(\varepsilon)\}$ gilt
 $|E_{i,P} a(x(\tau)) - E_{i,P} a(x(\tau'))| \leq \varepsilon$.

Die Ersetzung von \bar{X} durch die endliche Menge

$X(\varepsilon) = \{\tau : \tau \in \bar{X}, \tau \leq n(\varepsilon)\}$ verursacht also höchstens einen Fehler der Größe ε bei der Auszahlung. Die diskret gemischte Erwartung von $(X(\varepsilon), Y, K_i)$ ist definit und daraus folgert man die Definitheit von $\bar{\Gamma}_i^d$, da $\varepsilon > 0$ beliebig vorgegeben war.

Ist die Bedingung (2) erfüllt, so zeigt man, daß jede Folge

$$(P_n) \text{ bzgl. der Waldschen Metrik } \varphi_i(P, P') = \sup_{\bar{X}} |K_i(\tau, P) - K_i(\tau, P')|$$

$(P, P' \in Y)$ auf Y eine Cauchy-Teilfolge besitzt. Daraus kann man auf die Definitheit von $\bar{\Gamma}_i^d$ schließen (vergl. Bierlein [1], Wald [12]).

Dazu überlegt man sich zunächst, daß zu einer Folge (P_n) eine absorbierende Matrix P_0 und eine Teilfolge (P_{n_t}) existieren mit

$$\lim_{n_t} P_{n_t} = P_0 \text{ (elementweise)}, S(P_0) = S. \text{ Indem man gegebenenfalls}$$

noch einmal zu Teilfolgen übergeht, erhält man weiter
 $\lim_{n_t} G_{n_t}(j, k) = G_0(j, k) \quad (j, k \notin S) \text{ für } G_{n_t} = \sum_{s=0}^{\infty} P_{n_t}^s, \quad G_0 = \sum_{s=0}^{\infty} P_0^s.$

Daraus folgert man, daß zu $\varepsilon > 0$ ein $n(\varepsilon)$ und $n_1(\varepsilon)$ existieren mit $P_{i, n_t}(\mathcal{F} > n(\varepsilon)) \leq \varepsilon$ für $n_t \geq n_1(\varepsilon)$,

$P_{i, 0}(\mathcal{F} > n(\varepsilon)) \leq \varepsilon$. Unter Benutzung dieser Abschätzungen sowie

$\lim_{n_t} P_{n_t} = P_0$ (elementweise) rechnet man dann $\lim_{n_t} \sum_i (P_{n_t}, P_0) = 0$

nach.

Bemerkung: In der Bedingung (1) wird eine Gleichförmigkeit für den Eintrittszeitpunkt in S bei Start in $i \in E$ gefordert, in (2) eine Gleichförmigkeit des Rekurrenzverhaltens der Matrizen aus Y . Das folgende einfache Beispiel zeigt, daß die Bedingungen nicht notwendig sind. Es sei $E = \{1, \dots, N\}$, $S = \{N\}$, $Y = \{P: P \text{ absorbierend}, S(P) = \{N\}\}$, $a: E \rightarrow R^+ = \{y: y \geq 0\}$ mit $a(N) = 0$. Für $i \in E$ ist jedes Spiel $\Gamma(i)$ definit mit Spielwert $a(i)$, jedoch sind die Bedingungen (1) und (2) nicht erfüllt.

3. Eine Lösung für ein einfaches Stop-Problem kann man mit Hilfe der Dynamischen Optimierung gewinnen (vergl. [2], [3], [5]). Entsprechend der hier gegebenen Situation ändert man diese Verfahren ab und definiert für $f: E \rightarrow R$, beschränkt, $j \in E$:

$$\begin{aligned} (3.1) \quad V^0 f(j) &= f^+(j) \quad (= \max \{f(j), 0\}) \\ V^n f(j) &= \max \{V^{n-1} f(j), \inf_Y P V^{n-1} f(j)\} \\ U^0 f(j) &= f^+(j) \\ U^n f(j) &= \max \{f^+(j), \inf_Y P U^{n-1} f(j)\}. \end{aligned}$$

Analog wie bei einfachen Stop-Problemen erhält man

Lemma (3.2)

Es sei $\Gamma = (X, Y, K)$ ein Stop-Spiel mit $a(j) \geq 0$ ($j \in S$). Dann existiert $v(i) = \lim_n V^n a(i) = \lim_n U^n a(i)$ und es gilt

$$v(i) = \max \left\{ a^+(i), \inf_Y P v(i) \right\} \quad (i \in E).$$

Mit der in Lemma (3.2) definierten Funktion v hat man eine Schranke für den unteren Spielwert erhalten:

Satz (3.3)

Es sei $\Gamma = (X, Y, K)$ ein Stop-Spiel, $a(j) \geq 0$ ($j \in S$), $v(i) = \lim_n V^n a(i)$ ($i \in E$) und für $\varepsilon > 0$ sei $M(\varepsilon, v) = \{ j: v(j) - a(j) \leq \varepsilon \}$.

Dann gilt

$$(1) \quad v(i) \leq \sup_X \inf_Y K_i(\tau, P) \quad (i \in E)$$

und die Trefferzeit $\tau_\varepsilon = \inf \{ t: x(t) \in M(\varepsilon, v) \}$ von $M(\varepsilon, v)$ erfüllt

$$(2) \quad v(i) \leq E_{1,P}^t a(x(\tau_\varepsilon)) + \varepsilon \quad (P \in Y).$$

Beweis:

Man hat nur (2) zu zeigen, da (1) unmittelbar aus (2) folgt.

Zum Beweis von (2) stellt man die Funktion v in Form einer Riesz-Darstellung dar bzgl. der Matrix $(I_{E-M(\varepsilon, v)}^P)$, dabei sei $P \in Y$ und für $A \subseteq E$ sei $I_A(i, j) = \begin{cases} 1: & i = j \in A \\ 0: & \text{sonst} \end{cases}$

Mit Hilfe des Lemma (3.2) sowie $\lim_n (I_{E-M(\varepsilon, v)}^P)^n v(j) = 0$

($j \in E$) erhält man aus dieser Darstellung von v direkt die gewünschte Abschätzung.

Bei der Berechnung von v wird zeilenweise ein Infimum gebildet, das bedeutet, daß man Spieler 2 quasi zugesteht, zeilenweise eine für Spieler 1 ungünstige Strategie zu bestimmen. Man erhält daher dementsprechend

Korollar (3.4)

Es sei $\Gamma = (X, Y, K)$ ein Stop-Spiel, $a(j) \geq 0$ ($j \in S$), $Q = (P^{(k)}(k, j)) \in Y$, falls $P^k \in Y$ ($k \in E$) gilt. Dann ist $\Gamma'(i) = (X, Y, K_i)$ definit mit Spielwert $v(i)$ und die Trefferzeit τ_ϵ von $M(\epsilon, v)$ ist eine ϵ -gute Minimaxstrategie für Spieler 1 für $\epsilon > 0$, $i \in E$, falls eine der folgenden Bedingungen erfüllt ist:

(1) Es ist $a: E \rightarrow R^+$.

(2) Es existiert eine Konstante $C > 0$ mit

$$\sum_{t=0}^{\infty} P^t(j, k) \leq C \quad (k \notin S, P \in Y).$$

Beweis:

zu (1): Es sei (ϵ_n) eine positive fallende Nullfolge. Zu ϵ_n bestimmt man $P_n \in Y$ mit $P_n v(i) \leq \inf_Y P v(i) + \epsilon_n$. Man darf

annehmen, daß eine stochastische Matrix P_0 existiert mit $\lim_n P_n = P_0$ (elementweise). Für $v_n(i) = \sup_X K_i(\tau, P_n)$ ($n = 0, 1, \dots$) gilt dann $v_0(i) = \lim_n v_n(i)$. Dies folgert man aus Lemma 2 von Hoffman, Karp [10]. Denn die v_n ergeben sich als Lösungen Linearer Programme (vergl. [5]), unter den Voraussetzungen hier ist der Optimalwert der Programme stetig in den Koeffizienten. Andererseits hat man $v_0(i) \leq v(i) \leq v_n(i)$ ($i \in E$, $n = 1, 2, \dots$), also muß $\lim_n v_n(i) = v(i)$ gelten.

Zu (2): Analog wie in (1) bestimmt man zu (ϵ_n) eine Folge (P_n) .

Zu dieser Folge existiert eine Teilfolge (P_{n_t}) und eine stochastische Matrix P_0 , so daß gilt $\lim_{n_t} P_{n_t} = P_0$ (elementweise),

$\lim_{n_t} \sum_i (P_{n_t}, P_0) = 0$ und $P_0 v(i) = \inf_Y P v(i)$ ($i \in E$) (vergl. den

Beweis von Satz (2.2)). Für $v_{n_t}(i) = \sup_X K_i(\tau, P_{n_t})$ ergibt sich

dann: $0 \leq \lim_{n_t} [v_{n_t}(i) - v(i)] \leq \lim_{n_t} \sum_i (P_{n_t}, P_0) = 0$ ($i \in E$).

Inwieweit die Schranke $v(i)$ in praktischen Fällen numerisch bestimmt werden kann, hängt entsprechend (3.1) und Lemma (3.2) vom speziellen Aussehen von Y ab. Diese Berechnung ist insbesondere dann einfach, wenn Y endlich ist oder von einer endlichen Menge Y_0 erzeugt wird, das soll heißen

$$(*) Y = \{P : P = (P^{(k)}(k, j)), P^{(k)} \in Y_0 (k \in E)\}.$$

Unter diesen Voraussetzungen über Y gilt dann Lemma (3.2) sowie Satz (3.3) auch für nicht endliche, höchstens abzählbare Zustandsräume. Dementsprechend erhält man die Aussage von Korollar (3.4) für diese Zustandsräume, wenn die Bedingungen (1), (2) ersetzt werden durch (*). (vergl. auch [7]).

Literatur

- [1] D. Bierlein: Spieltheorie-
Ausarbeitung einer Vorlesung an der Universität Karlsruhe im WS 1967/68, als Manuskript vervielfältigt.
- [2] L. Breiman: Stopping-Rule Problems-
Applied Combinatorial Mathematics (Hrsg.: E. F. Beckenbach), New York, London, Sidney 1964, S. 284 - 319.

- [3] C. Derman: Finite State Markovian Decision Processes-
New York, London 1970.
- [4] E.B.Dynkin: Optimal Selection of Stopping Time for a
Markov Process-
Dokl.Akad.Nauk USSR 150, 1963, S. 238-240.
- [5] E.B.Dynkin, Sätze und Aufgaben über Markoffsche Prozesse:
A.A.Juschkevitsch: Heidelberg 1969.
- [6] B.H. Goldstein: Potentialtheorie Markoffscher Ketten-
Op.Res.Verf. IV (Hrsg.: R. Henn), Meisenheim
1967, S. 271-594.
- [7] B.H. Goldstein: Über Stop-Probleme mit alternativen Aus-
zahlungsfunktionen und Übergangswahr-
scheinlichkeiten-
im Druck in: Op.Res.-Verfahren, Bd. 12, 1970
Ergebnisband der IV. Oberwolfach-Tagung über
Operations Research (Hrsg.: R. Henn, H.P.Künzi
H. Schubert)
- [8] R. Henn, Einführung in die Unternehmensforschung I,II
H.P. Künzi: Berlin, Heidelberg, New York 1968.
- [9] K.H. Hinderer: Foundations of Non-stationary Dynamic
Programming with Discrete Time Parameter-
Lecture Notes in Operations Research and
Mathematical Systems, (Hrsg.: M. Beckmann,
H.P. Künzi), Bd. 33, 1970.
- [10] J. Hoffman, On Non-terminating Stochastic Games-
R.M.Karp: Man.Sci. 12, 1966, S. 359-370.
- [11] J.G. Kemeny, Finite Markov Chains-
J.L. Snell: Princeton, N.J., 1963.
- [12] A. Wald: Statistical Decision Functions-
New York, 1960.

Privatdozent Dr. Bernd H.Goldst
Institut für Statistik und
Quantitative Methoden der
Unternehmensführung
Universität (TH) Karlsruhe
75 Karlsruhe
Kaiserstraße 12

Classifying the States of a Finite Markov Chain

by G. Uebe, München

ABSTRACT

In network flow theory the most efficient way to determine the shortest path is to apply a triple operation $d_{ij} = \text{Min}[d_{ij}, d_{ik} + d_{kj}]$ on the distance matrix $((d_{ij}))$. A completely analogous triple operation $b_{ij} = \text{Max}[b_{ij}, b_{ik}b_{kj}]$ can be defined in a Boolean matrix $((b_{ij}))$, $b_{ij} = 0, 1$. Applying this operation, **all connections between points i and j are uncovered.**

Defining $b_{ij} = 0$ or 1 , depending on whether the transition probability matrix of a finite Markov chain has $p_{ij} = 0$ or $\neq 0$, all cells with nonzero higher transition probabilities are determined. An upper bound of the number of matrix multiplications, to obtain them, follows immediately.

Using the criterion state i and state j are equivalent (belong to the same class) if $b_{ij} = b_{ji} = 1$ the equivalent states can be collected easily.

Ordering the equivalent set of states one obtains a block triangular matrix of persistent sets of states and transient sets of states. The transiency follows from the criterion $\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$, which is immediately applicable after the reordering.

1. The classification problem

Given a finite Markov chain with n states S_1, S_2, \dots, S_n and the transition matrix $((p_{ij}))$ $i = 1, 2, \dots, n$, $j = 1, 2, \dots, n$ it is useful to classify the states [Feller, pp. 349-357]. Using a triple operation

$$b_{ij} := \max[b_{ij}, \max(b_{ik} \cdot b_{kj})]$$

on the Boolean matrix $((b_{ij}))$ with $b_{ij} = 0$ or 1 and

$$b_{ij} = 1 \quad \text{if} \quad p_{ij} > 0$$

$$0 \quad \text{if} \quad p_{ij} = 0$$

one can easily establish whether a Markov chain is irreducible or not. If it is reducible the well known criterion

$$[\text{A state } S_j \text{ is transient}] \iff \sum_{\ell=1}^{\infty} p_{jj}^{(\ell)} < \infty$$

can be applied, where $p_{jj}^{(\ell)}$ is the higher transition probability, in general

$$p_{ij}^{(\ell)} = \sum_{k=1}^n p_{ik}^{(\ell-1)} p_{kj} = \sum_{k=1}^n p_{ik} p_{kj}^{(\ell-1)}$$

$$i = 1, 2, \dots, n$$

$$j = 1, 2, \dots, n$$

Before presenting constructive criteria

- (i) of the reducibility of a Markov chain
- (ii) of the transiency of the states of a Markov chain

a number of definitions and properties of Markov chains taken from Feller are reproduced.

2. Definitions and properties used

- (1) A set C of states is irreducible if and only if every state of the set can be reached from every other state of the set
- (2) A set C of states is closed if no state outside C can be reached from any state in C
- (3) A set C of states is a minimal closed set if it is closed and irreducible
- (4) A Markov chain which lacks property (1), i.e. a reducible Markov chain, can be reordered by simultaneous permutation of rows and columns into a block diagonal matrix, in which the blocks on the diagonal correspond to minimal closed sets.

Example

	1	2	3	4	5
1	1/2			1/2	
2		1/4			3/4
3	1/2	1/8	1/8	1/8	1/8
4	2/3			1/3	
5		1/5			4/5

before ordering

	1	4	2	5	3
1	1/2	1/2			
4	2/3	1/3			
2			1/4	3/4	
5			1/5	4/5	
3	1/2	1/8	1/8	1/8	1/8

after ordering

There are the closed sets $C_1 = (1, 4)$, $C_2 = (2, 5)$, $C_3 = (1, 4, 2, 5)$ and $C_4 = (1, 2, 3, 4, 5)$, while minimal closed sets are only C_1 and C_2 .

In order to find all closed and minimal closed sets, it suffices to know which p_{ij} are zero and which are positive. This gives a Boolean matrix B

$$B = ((b_{ij}))$$

where

$$b_{ij} = \begin{cases} 1 & \text{if } p_{ij} > 0 \\ 0 & \text{if } p_{ij} = 0 \end{cases} \quad (1)$$

For the example one obtains

1			1	
	1			1
1	1	1	1	1
1			1	
	1			1

Obviously in general these matrices B are not so small and nice, in the sense that all possible connections are already available and such that one can recognize the ordering by inspection. Using higher transition probabilities $((p_{ij}^{(k)}))$, some initial coefficients equal to zero may become positive. One way to find that out is to multiply P sufficiently often. But this is inefficient. Using the triple operation

$$b_{ij} := \max [b_{ij}, \max(b_{ik} \cdot b_{kj})] \quad (2)$$

on matrix B all possible connections are found much faster. This is a direct borrowing from network flow theory, from the shortest path algorithm [see e. g. T. C. Hu, p. 154-161].

3. Example [Feller, p. 350]

	1	2	3	4	5	6	7	8	9
1				1					1
2		1	1		1				1
3								1	
4	1								
5					1				
6		1							
7		1				1	1		
8			1						
9				1					1

Applying operation (2) gives

	1	2	3	4	5	6	7	8	9
1	1			1					1
2	1	1	1	1	1			1	1
3			1					1	
4	1			1					1
5					1				
6	1	1	1	1	1			1	1
7	1	1	1	1	1	1	1	1	1
8			1					1	
9	1			1					1

Keeping track of the number of replacements in each entry the largest of these gives an upper bound on the number of multiplications of B , necessary to produce all connections, say k , i.e. with P^k all transitions have been covered at least once. For the example $k = 5$, but the number of operations is far less than those of five matrix multiplications due to (2). If C is any set; e.g. an irreducible set, a closed set, or a minimal closed set, and if S_i and S_j are two states, and if $S_i \in C$ and if

$$b_{ij} = b_{ji} = 1 \quad (3)$$

which obviously implies that $b_{ii} = 1$, then S_j must belong to the same set C , i.e.

$$[b_{ij} = b_{ji} = 1] \iff [S_i \text{ and } S_j \text{ belong to one set}] \quad (4)$$

As can be seen from simple examples, this aggregation does not imply properties of the sets per se, i.e. aggregation can take place within minimal closed sets (figure 1), within closed, but reducible sets (figure 2) and within sets which are not closed (figure 3). The pair to be aggregated is circled.

	1	2	3
1	1/2	1/2	
2	1/2	1/2	
3			1

$$C = \{1, 2\}$$

C is minimal closed

FIGURE 1

	1	2	3
1	1/2		1/2
2		1	
3	1/2		1/2

$$C = \{1, 2, 3\}$$

C is closed, but reducible

FIGURE 2

	1	2	3
1	1		
2	1/3	1/3	1/3
3	1/3	1/3	1/3

$$C = \{2, 3\}$$

C is not closed

FIGURE 3

Beginning the aggregation of states we do not know which set of states will finally result, and obviously an aggregation rule like "Aggregate those S_j , for which (3) is true, into S_i " is not unique; one could aggregate S_i into S_j equally well. But equally obvious is that any such rule may only change the name of the set, its membership is uniquely determined by (4).

Applying the aggregation to the example of page 5 gives three minimal closed sets

$$C_1 = \{5\}$$

$$C_2 = \{3, 8\}$$

$$C_3 = \{1, 4, 9\}.$$

Each of these minimal closed sets is a stochastic submatrix and their order does not matter, i.e. a sequence may be

$C_1 \quad C_2 \quad C_3$

or

$C_2 \quad C_3 \quad C_1 \quad \text{etc.}$

Result is e.g. the following reordered incidence matrix B

C_1

C_2

C_3

5

3

8

1

4

9

2

6

7

1'

2'

3'

4'

5'

6'

7'

8'

9'

5	1'	x								
3	2'		x	x						
8	3'		x	x						
1	4'				x	x	x			
4	5'				x	x	x			
9	6'				x	x	x			
2	7'	x	x	x	x	x	x	x		
6	8'	x	x	x	x	x	x	x		
7	9'	x	x	x	x	x	x	x	x	

C_1

C_2

C_3

4. The aggregation of the Markov Chain

The remaining square matrix R of the canonical form of the Markov chain [Kemeny-Snell, p. 36]

$$P = \begin{pmatrix} C & 0 \\ N & R \end{pmatrix} \tag{5}$$

where C corresponds to C, the set of minimal closed sets, and R corresponds to

R the set of remaining states, if there are any, and N is a matrix which is not a zero matrix, is of particular structure.

Due to the aggregation R is changed into a matrix \tilde{B} of size $r \times r$ ($r < n$), in which there is no pair of states S_k and S_ℓ $k = 1, 2, \dots, r$, $\ell = 1, 2, \dots, r$, such that

$$\tilde{b}_{k\ell} = \tilde{b}_{\ell k} = 1 \quad \ell \neq k \quad (6)$$

is true.

1. Denoting by

$\tilde{b}_{k.}$ the sum of ones in row k of \tilde{B} , and

$\tilde{b}_{.\ell}$ the sum of ones in column ℓ of \tilde{B} ,

then for each row and column k of \tilde{B} we obtain

$$\tilde{b}_{k.} + \tilde{b}_{.k} \leq r+1 \quad \text{for } k = 1, 2, \dots, r. \quad (7)$$

Therefore without loss of generality rows and columns can be permuted simultaneously such that

$$\tilde{b}_{1.} \leq \tilde{b}_{2.} \leq \tilde{b}_{3.} \dots \leq \tilde{b}_{r.} \quad (8)$$

and

$$\tilde{b}_{.1} \geq \tilde{b}_{.2} \geq \tilde{b}_{.3} \dots \geq \tilde{b}_{.r}, \quad (9)$$

where

$$\tilde{b}_{.k} \equiv r+1 - \tilde{b}_{k.} \geq \tilde{b}_{.k} \quad \text{for } k = 1, 2, \dots, r \quad (10)$$

If (7) is satisfied as an equality for all k , then (10) is satisfied as an equality for all k and in (9) $\tilde{b}_{.k}$ can be replaced by $\tilde{b}_{k.}$. But as soon as this strong requirement of equality [see Ryser, p. 110, for such a case of tournament matrices] is weakened to the more general statement of (7), which permits strict inequality, then the row ordering (8) does not impose a corresponding monotonic column ordering. The following example illustrates this point.

		columns k				
rows k		1	2	3	4	$b_{k.}$
	1	1				1
	2	1	1			2
	3		1	1		2
	4		1	1	1	3
	$b_{.k}$	2	3	2	1	

Obviously (8) is satisfied, and implies (9), namely

$$1 \leq 2 \leq 2 \leq 3 \quad (8')$$

$$4 \geq 3 \geq 3 \geq 2. \quad (9')$$

But replacing $\tilde{b}_{.k}$ by $\tilde{b}_{k.}$ in (9)' does not work

$$2 \not\geq 3 \geq 2 \geq 1.$$

But (8) suffices to prove the next theorem.

2. Theorem. The matrix \tilde{B} can be re-arranged by simultaneous permutations of rows and columns so as to be triangular, i.e. there are no entries to the right of the main diagonal of the re-ordered matrix $\tilde{\tilde{B}}$, and \tilde{B} , respectively $\tilde{\tilde{B}}$, has at most $\frac{r(r+1)}{2}$ positive elements.

Proof: Consider the first row of \tilde{B} , which is ordered according to (8). Assume there is a positive entry off the diagonal, i.e. $\tilde{b}_{1k} = 1$, $k \neq 1$.

Since $\tilde{b}_{kl} = 0$ by assumption and $\tilde{b}_{k.} \geq \tilde{b}_{1.}$ by (8), there must be at least one positive entry $\tilde{b}_{k\ell} = 1$, $\ell \neq k \neq 1$ such that $\tilde{b}_{1\ell} = 0$. But this is a contradiction to the fact that the triple operation has been concluded: we could produce an additional entry $\tilde{b}_{1\ell} := \tilde{b}_{1k} \cdot \tilde{b}_{k\ell} = 1$. Hence there is no off diagonal element in the first row, and if there is a positive entry it must be on the main diagonal. So the first state of \tilde{B} is fixed. Deleting the first row and the first column of \tilde{B} , we obtain a new matrix \tilde{B}_1 , the dimension of which is reduced by one and which has row-column-sum properties corresponding to (7). Assuming that a row ordering with a simultaneous column ordering has been performed in \tilde{B}_1 , which corresponds to (8), then the arguments applying to the first row of \tilde{B} , apply equally to the first row of \tilde{B}_1 , and the state corresponding to the first row of \tilde{B}_1 is fixed as the second state of \tilde{B} . Deleting the first row and the first column of \tilde{B}_1 , a matrix \tilde{B}_2 is obtained. After re-arrangement of \tilde{B}_2 according to (8) the above argument of the non-existence of an off-diagonal positive element in the first row applies again. Continuing with $\tilde{B}_3, \tilde{B}_4, \dots, \tilde{B}_{r-1}$ finally triangularity is obtained. A triangular matrix obviously has at most $\frac{r(r+1)}{2}$ positive elements. Q.E.D.

This concludes the problem of reducibility of the Markov chain. Before using this partitioning of the states for classification, some results are needed [Feller 352-353].

5. The classification by the recurrence time criterion

Each state S_j is characterized by its recurrence time distribution $\{f_j^{(k)}\}$, where $f_j^{(k)}$ is the probability that the first return to S_j occurs at time k

$$\begin{aligned}
 f_j^{(1)} &= p_{jj} \\
 f_j^{(2)} &= p_{jj}^{(2)} - f_j^{(1)} p_{jj} \\
 &\vdots \\
 f_j^{(k)} &= p_{jj}^{(k)} - \sum_{\ell=1}^{k-1} f_j^{(\ell)} p_{jj}^{(k-\ell)} \quad k \geq 2.
 \end{aligned} \tag{11}$$

Adding all $f_j^{(k)}$, gives the probability f_j that starting from S_j there is a return to S_j

$$f_j = \sum_{k=1}^{\infty} f_j^{(k)} \tag{12}$$

Now the classification of the states can be made; states are either transient or persistent.

$$\text{A state } S_j \text{ is transient if } f_j < 1 \tag{13a}$$

and

$$\text{a state } S_j \text{ is persistent if } f_j = 1 \tag{13b}$$

A necessary and sufficient condition for $f_j < 1$ is

$$\sum_{k=1}^{\infty} p_{jj}^{(k)} < \infty \tag{14}$$

[For a proof see Feller]

Criterion (14) is usually difficult to apply [see Feller's remark, p. 354], but having ordered the Markov chain, the application is immediate.

Theorem: (i) The minimal closed sets. the stochastic submatrices are persistent.

If the submatrix is of size 1×1 , it is called an absorbing state.

(ii) The states not contained in the minimal closed sets are transient states.

Proof: (i) is obvious, since by construction for any of the stochastic submatrices all entries are positive, (not necessarily simultaneously, if the submatrix is not regular) i.e. after multiplying the submatrix sufficiently often, but finitely often, all states are connected and by definition the minimal closed set is never left. For all j in the closed set f_j must be 1.

(ii) Follows from the definition of matrix multiplication for triangular matrices. Aggregating the original matrix P (5), in its canonical form, one obtains a matrix \tilde{P}

$$\tilde{P} = \begin{pmatrix} \overbrace{I}^s & \overbrace{0}^r \\ \underbrace{\tilde{N}}_s & \underbrace{\tilde{B}}_r \end{pmatrix} \quad (15)$$

where the identity matrix I of size $s \times s$ represents the s sets of persistent states and the triangular matrix \tilde{B} of size $r \times r$ represents the r sets of transient states ($r + s \leq n$).

Considering the multiplication of the ordered Markov-chain (15), all powers of \tilde{P} have again this partitioning, and the elements on the main diagonal of \tilde{B} are powers of its diagonal elements. But by construction the rows pertaining to \tilde{B} have coefficients > 0 outside \tilde{B} , i.e., all diagonal elements in \tilde{B} are strictly smaller than one

$$0 \leq \tilde{p}_{jj} < 1 \quad \text{for } j \in \mathcal{B} \quad (16)$$

$$\tilde{p}_{jj}^{(k)} = \tilde{p}_{jj}^k \quad \text{for } j \in \mathcal{A} \quad (17)$$

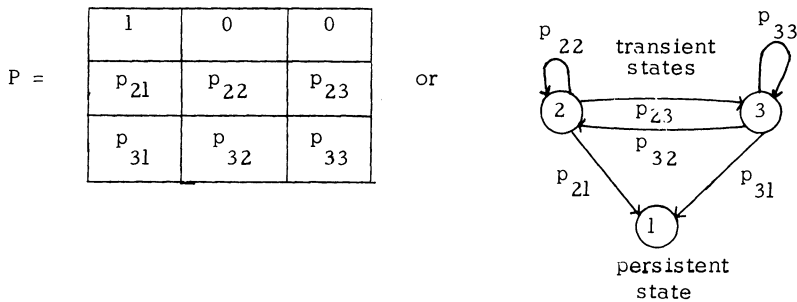
Therefore

$$\sum_{l=1}^{\infty} \tilde{p}_{jj}^{(l)} = \sum_{l=1}^{\infty} \tilde{p}_{jj}^l = \frac{\tilde{p}_{jj}}{1 - \tilde{p}_{jj}} < \infty \quad (18)$$

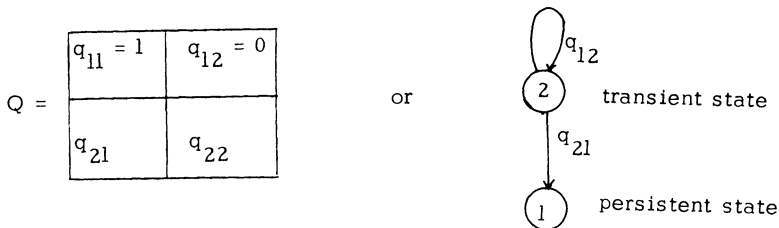
and all states in \mathcal{B} are transient.

6. The case of non-lumpable chains

The actual value of \tilde{p}_{jj} ($j \in \mathcal{A}$) is of no interest since one can always choose a majorizing value, which satisfies (16) and thus suffices for (18). This is important since in general the states of a Markov chain do not need to satisfy lumpability conditions [Kemeny-Snell, p. 123-140], as may be demonstrated by an example of one absorbing state and two transient states:



Aggregating state 2 and state 3, as proposed, one considers a Markov chain



Considering a distribution $(p_i(t))$ $i = 1, 2, 3$ of the original states and the distribution $(\pi_k(t))$ $k = 1, 2$ of the aggregated states both must produce the same distribution for the next period, i. e.

$$\pi_1(t+1) = p_1(t+1) \tag{19}$$

$$\pi_2(t+1) = p_2(t+1) + p_3(t+1)$$

and

$$\pi_1(t+1) = \pi_1(t)q_{11} + \pi_2(t)q_{21} \quad (20)$$

$$\pi_2(t+1) = \pi_1(t)q_{12} + \pi_2(t)q_{22}$$

and

$$p_1(t+1) = p_1(t)p_{11} + p_2(t)p_{21} + p_3(t)p_{31} \quad (21)$$

$$p_2(t+1) = p_1(t)p_{12} + p_2(t)p_{22} + p_3(t)p_{32}$$

$$p_3(t+1) = p_1(t)p_{13} + p_2(t)p_{23} + p_3(t)p_{33}$$

By comparison of terms

$p_{11} = 1$	$p_{12} + p_{13} = 0$
$Q = \frac{p_2(t)p_{21} + p_3(t)p_{31}}{p_2(t) + p_3(t)}$	$\frac{p_2(t)[p_{22} + p_{23}] + p_3(t)[p_{32} + p_{33}]}{p_2(t) + p_3(t)}$

Assuming that at least either $p_2(t) > 0$ or $p_3(t) > 0$ the time dependence of $q_{22} = q_{22}(t)$ does not affect (18), i.e., one can choose a \tilde{p}_{22} such that $q_{22}(t) \leq \tilde{p}_{22}$ and \tilde{p}_{22} satisfies (18).

7. The steps of the classification

Returning to the original matrix P , the criterion for whether a set is a persistent one or a transient one, is therefore either

$$\sum_{j \in C_k} p_{ij} = 1 \text{ for all } i \in C_k \quad (22)$$

$k = 1, 2, \dots, s$ the number of sets of persistent states

or

$$\sum_{j \in \mathcal{R}_k} p_{ij} < 1 \quad \text{for all } i \in \mathcal{R}_k \quad (23)$$

$k = 1, 2, \dots, r$ the number of sets of transient states

where $(\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s)$ and $(\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_r)$ are the sets of states corresponding to C and R respectively (5).

Summarizing, classifying a Markov chain can be done as follows

Step 1. For the Markov chain matrix P a Boolean matrix B is defined.

Step 2. Applying the triple operation on B all connections in P are revealed.

Step 3. Aggregating equivalent states, $s + r$ sets of states are determined.

Step 4. Reordering P according to these sets, one obtains s sets of persistent states and r sets of transient states.

The author wishes to thank Professor J. B. Rosser for a very helpful discussion of the paper.

REFERENCES

1. J. Eufinger, Operations on directed graphs, Crelle's Journal für die reine und angewandte Mathematik, 1970.
2. William Feller, An Introduction to Probability Theory and Its Applications, 2nd ed., John Wiley and Sons, New York, 1957.
3. T. C. Hu, Integer Programming and Network Flows, Addison Wesley, Reading, Mass. 1969.
4. J. G. Kemeny and J. L. Snell, Finite Markov Chains, D. van Nostrand Co., Princeton, New Jersey, 1960.
5. H. J. Ryser, Matrices of Zeros and Ones in Combinatorial Mathematics, in "Recent Advances in Matrix Theory," Hans Schneider, editor, The University of Wisconsin Press, Madison and Milwaukee, 1964, pp. 103-124.
6. A. Jaeger, K. Wenke, Lineare Wirtschaftsalgebra, Vol. 2, Teubner, Stuttgart, 1969.

Optimales Stoppen von endlichen Markoff-Ketten

von O. Emrich, Augsburg

- I. Zur Einführung in die Problemstellung soll das folgende einfache Beispiel betrachtet werden.

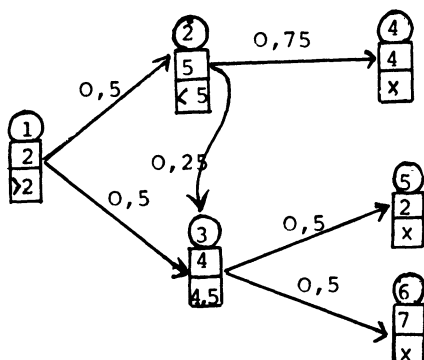
Ein Grundstück in einem verkehrstechnisch unerschlossenen Gebiet ist zur Zeit zwei Geldeinheiten, etwa (DM 200.000,--) wert und kann auch zu diesem Preis verkauft werden. Zur Verbesserung der Verkehrslage sind zwei Pläne in der Diskussion, die beide die gleiche Chance haben realisiert zu werden: Bau einer Landstraße und Bau einer Hauptdurchgangsstraße. Der Bau einer Landstraße erhöht den Grundstückswert auf fünf Einheiten, der Bau der Durchgangsstraße wegen des Lärms nur auf vier Einheiten. Nach Fertigstellung der Landstraße bestehen die Möglichkeiten:

- a) Ausbau zur Durchgangsstraße (Realisierungswahrscheinlichkeit $\frac{1}{4}$) oder
- b) Errichtung einer Fabrik in der Nähe mit starker Umweltverschmutzung (Realisierungswahrscheinlichkeit $\frac{3}{4}$; Wertminderung auf vier Einheiten).

Nach Fertigstellung der Durchgangsstraße sind zwei Entwicklungen möglich:

- a) Bau eines Flughafens so, daß das Grundstück in der Einflugschneise liegt (Realisierungswahrscheinlichkeit $\frac{1}{2}$; Wertminderung auf zwei Einheiten);
- b) Errichtung einer Wohnsiedlung (Realisierungswahrscheinlichkeit $\frac{1}{2}$; Wertsteigerung auf sieben Einheiten).

Das Problem besteht nun darin, den "optimalen" Verkaufszeitpunkt festzustellen.



- Nummer der Ausbaustufe
- Augenblicklicher Verkaufswert
- ◁ erwarteter Verkaufswert, wenn nicht verkauft wird

Bemerkung: in den Ausbaustufen

④, ⑤, ⑥ muß verkauft werden.

Als optimal soll ein Verhalten bezeichnet werden, wenn der zu erwartende Verkaufspreis maximal wird (vgl. auch [7]).

II. Diese Problemstellung^{*)} wird durch das folgende Modell erfaßt:

Man betrachtet:

- a) eine homogene, zeitdiskrete Markoff-Kette $(X_t)_{t \in \mathbb{N}}$ mit dem Zustandsraum $Z = \{1, 2, 3, \dots, z\}$ der Übergangsmatrix $P = (p(i, j))_{i, j \in Z}$ und der Anfangsverteilung m auf Z

- b) eine Auszahlungsfunktion $a : Z \rightarrow \mathbb{R}^+$

^{*)} Weitere Anwendungsmöglichkeiten werden in [6] diskutiert.

wobei die folgenden "Spielregeln" gelten:

- c) beim Abbrechen im Zustand i wird $a(i)$ ausbezahlt; es ist eine Stoppzeit τ so zu bestimmen, daß die zu erwartende Auszahlung maximal wird, d.h., so daß für jede andere Stoppzeit τ' gilt:

$$(1) \quad E a(X_{\tau},) \leq E a(X_{\tau'}) \quad \text{für alle } \tau'$$

Definition:

Eine Stoppzeit τ die (1) erfüllt, heißt optimale Stoppzeit.

Bei endlichen Zustandsräumen (und nur diese sollen hier untersucht werden) ist nun eine einfache Charakterisierung der optimalen Stoppzeiten möglich:

Satz 1:

Ist der Zustandsraum Z endlich, dann existiert eine Teilmenge $S \subseteq Z$, so daß

$$(2) \quad \tau_S := \inf \{t : X_t \in S\}$$

eine optimale Stoppzeit ist, unabhängig von der Anfangsverteilung m .

Bemerkung: Der Beweis von Satz 1 ist z. B. in [3] zu finden.

Folgerung 2. Satz 1 sichert auch die Existenz optimaler Stoppzeiten.

Folgerung 3. Ein Stopp-Problem ist durch die Angabe von (Z, P, a) vollkommen beschrieben.

Definition:

S heißt optimale Stoppmenge, falls das nach (2) bestimmte τ_S eine optimale Stoppzeit ist.

Definition:

$v : Z \rightarrow \mathbb{R}^+$ heißt Spielwert des Stopp-Problems
 (Z, P, a) falls

$$v(i) := \sup \{E(a(X_\tau) | X_0 = i); \tau \text{ Stoppzeit}\}$$

Satz 4:

Der Spielwert v genügt der folgenden Beziehung:

$$(3) \quad v(i) = \max \{a(i), \sum_{j \in Z} p(i, j) v(j)\}$$

Bemerkung: Der Beweis ist z.B. in [3] zu finden.

Folgerung 5:

$$(4) \quad v(i) \geq a(i)$$

$$(5) \quad v(i) \geq \sum_{j \in Z} p(i, j) v(j) = Pv(i)$$

Folgerung 6:

Ist v der Spielwert, dann ist die Stützmenge

$$S^* := \{i : v(i) = a(i), j \in Z\}$$

eine optimale Stoppmenge.

Bemerkung: Durch eine der Angaben v , S oder τ ist damit das Stopp-Problem gelöst.

Definition:

Genügt f der Bedingung (5) (d.h. $f \geq Pf$), so heißt f Supermartingal. ($f : Z \rightarrow \mathbb{R}$)

Die für das folgende entscheidende Aussage bringt

Satz 7:

Der Spielwert v ist das kleinste a majorisierende Supermartingal.

Bemerkung: Der Beweis von Satz 7 ist z.B. in [3] zu finden.

Folgerung 8:

Der Spielwert v ist die kleinste unter den Funktionen, die (4) und (5) erfüllen; d.h. genügt f der Forderung

$$(6) \quad \sum_{i \in Z} f(i) = \text{Min } !$$

unter den Nebenbedingungen (4), (5), so ist f gleich dem Spielwert.

Bemerkung: Der Spielwert v ist Lösung der durch (4), (5) und (6) definierten linearen Optimierungsaufgabe.

Bemerkung: Setzt man $v_0 = a$ und ändert man (3) in

$$(3') \quad v_{k+1}(i) = \max \{a(i), \sum_{j \in Z} p(i,j) v_k(j)\}$$

so gilt $\lim_{k \rightarrow \infty} v_k = v$. (Verfahren der sukzessiven Approximation).

Mit den beiden letzten Bemerkungen sind zwei der bisher bekannten Lösungsmethoden angedeutet.

Eine weitere Lösungsmethode (die man oft auch mit dem Verfahren der sukzessiven Approximation zur Dynamischen Optimierung rechnet) sind die Verfahren der Politikiteration (vgl. [2]). Ein spezielles Verfahren der Politikiteration soll nun hier entwickelt werden.

III. Durch die folgenden Rekursionsformeln (a), (b), (c) werden Teilmengen $S_n \subset Z$ und Funktionen $w_n : Z \rightarrow \mathbb{R}^+$ definiert:

$$(a) \quad S_0 = \{i: i \in Z, p(i,i)=1\} \cup \{i: i \in Z, a(i)=\max \{a(j); j \in Z\}\}$$

$$w_0(i) = E(a(X_{\tau_{S_0}}) | X_0 = i) \quad \text{für alle } i \in Z$$

$$(b) \quad S_{n+1} = S_n \cup \{i: i \in Z, a(i) - w_n(i) = \max \{a(j) - w_n(j); j \in Z\}\}$$

$$(c) \quad w_{n+1}(i) = E(a(X_{\tau_{S_{n+1}}}) | X_0 = i)$$

Der folgende Satz sichert nun, daß durch (a), (b), (c) eine optimale Stoppmenge und der Spielwert bestimmt werden können (Verfahren der Stoppmengenerweiterung).

Satz 9:

Für $n \geq z$ gilt $w_n \geq a$. Ist $w_n \geq a$ so ist S_n eine optimale Stoppmenge und es gilt $w_n = v$.

Beweis: Zuerst zeigt man durch Induktion nach n (und zwar zunächst für absorbierende Markoff-Ketten, d.h. für Markoff-Ketten mit der Eigenschaft, daß von jedem Zustand i ein endlicher Weg positiver Wahrscheinlichkeit zu einem absorbierenden Zustand j , (d.h. $p(j,j) = 1$) führt): die Mengen S_n sind in der Stützmenge S^* enthalten und die w_n sind Supermartingale. Daß S_0 in der Stützmenge S^* enthalten ist, ist unmittelbar einsichtig (vgl. auch [3]). Die Supermartingaleigenschaft von w_0 ist leicht explizit nachzuweisen, wenn man bedenkt, daß w_0 ein Martingal zu der beim ersten Eintreten in S_0 gestoppten Markoff-Kette ist.

Unter der Voraussetzung: $S_n \subseteq S^*$ und w_n ist Supermartingal, ist nun zu zeigen: $S_{n+1} \subseteq S^*$ und w_{n+1} ist Supermartingal.

Das erste zeigt man, indem man nachweist $g := v - w_n$ ist ein Submartingal bezüglich der in S^* gestoppten Markoff-Kette; eine einfache Modifikation eines Satzes von Bauer [1] besagt dann, daß g sein Maximum auf den absorbierenden Zuständen, d.h. auf S^* annimmt; damit gelingt es nachzuweisen, daß $v(i) = a(i)$ für alle Zustände $i \in S_{n+1} \setminus S_n$; d.h. $i \in S^*$.

Die Supermartingaleigenschaft von w_{n+1} prüft man dann getrennt für die Zustände $i \in S_{n+1}$ und $i \in S_n$ nach. Mit der (trivialen) ersten Aussage von Satz 9 und Satz 7 folgt

$$w_z \geq v.$$

Andererseits ist $w_z(i) = E(a(X_{\tau_{S_z}}) | X_0 = i)$, d.h.

nach Definition des Spielwertes $v : v \geq w_z$.

Nun bleibt noch zu zeigen, daß die Voraussetzung einer absorbierenden Kette keine Einschränkung bedeutet. Dies gelingt mit Sätzen von Bauer [1], die besagen, daß die Menge aller Supermartingale die Zustände von minimalen Absorptionsmengen nicht trennen.

(Ein ausführlicher Beweis ist im Anhang angegeben.)

Die spezielle Art des obigen Algorithmus (a), (b), (c), der von einer kleinen Teilmenge S_0 des Zustandsraumes Z ausging und diese Menge in jedem Schritt um mindestens einen Zustand erweitert, bis die optimale Stoppmenge erreicht war, legt nun nahe, ein "komplementäres" Verfahren anzusetzen, das von dem gesamten Zustandsraum Z als Anfangsstoppmenge ausgehend die jeweiligen Stoppmengen S_1 solange verkleinert, bis eine optimale Stoppmenge erreicht ist.

Ein solches Verfahren der Stoppmengeneinschränkung wird durch die folgenden Beziehungen angegeben:

$$(a') \quad S_0 = Z, \quad w_0 = a$$

$$(b') \quad \bar{S}_{n+1} = \bar{S}_n \cup \{i : i \in S_n, Pw_n(i) > w_n(i)\} \quad 1)$$

$$(c') \quad w_{n+1}(i) = E(a(X_{\tau_{S_{n+1}}}) | X_0 = i)$$

Satz 10:

Es gilt $S_{z+1} = S_z$. Ist $S_{n+1} = S_n$, so ist S_n eine optimale Stoppmenge und $w_n = v$.

Der Algorithmus der Stoppmengeneinschränkung hat gegenüber dem der Stoppmengenerweiterung einige Nachteile, so läßt sich z.B. keine (nicht triviale) Abschätzung über die "Güte" der jeweiligen Stoppmenge S_n angeben im Gegensatz zum Verfahren der Stoppmengenerweiterung; dort gilt nämlich:

1) \bar{S} ist das Komplement von S , d.h. $\bar{S} = Z \setminus S$.

$$(6) \quad v(j) - w_n(j) \leq \max\{a(i) - w_n(i); i \in Z\} \quad \text{für } j \in Z$$

((6) folgt aus der Supermartingaleigenschaft von w_n).

IV. Ein weiterer wesentlicher Nachteil ist die geringere Verallgemeinerungsfähigkeit, so scheint eine direkte Bestimmung von sogenannten ϵ -guten Stoppmengen **nicht möglich zu sein.**

Definition:

Eine Menge $S \subseteq Z$ heißt ϵ -gute Stoppmenge (oder kurz ϵ -Stoppmenge), falls für den Spielwert v gilt:

$$E(a(X_{\tau_S}) | X_0 = i) \geq v(i) - \epsilon \quad \text{für alle } i \in Z \quad (\epsilon > 0)$$

Bestimmt man nun nach den folgenden Rekursionsformeln S_n bzw. w_n (ϵ -Verfahren der Stoppmengenerweiterung)

$$(a\epsilon) \quad S_0 = \{i: i \in Z, a(i) - \epsilon \geq \max\{a(j); j \in Z\}\}$$

$$a_0(i) = \begin{cases} a(i) & \text{für } i \notin S_0 \\ \max\{a(j); j \in Z\} - \epsilon & \text{für } i \in S_0 \end{cases}$$

$$w_0(i) = E(a(X_{\tau_{S_0}}) | X_0 = i)$$

$$(b\epsilon) \quad S_{n+1} = S_n \cup \{i: i \in Z, a(i) - w_n(i) - \epsilon \geq \max\{a(j) - w_n(j); j \in Z\}\}$$

$$a_{n+1}(i) = \begin{cases} a_n(i) & \text{für } i \in S_{n+1} \setminus S_n \\ \max\{a(j) - w_n(j); j \in Z\} + w_n(i) - \epsilon & \text{sonst} \end{cases}$$

$$(c\epsilon) \quad w_{n+1}(i) = E(a_{n+1}(X_{\tau_{S_{n+1}}}) | X_0 = i)$$

dann gilt der folgende

Satz 11:

Nach höchstens $K = \min\{z, \left\lceil \frac{\sup\{a(i); i \in Z\}}{\epsilon} \right\rceil\}$ Schritten

gilt: $w_K = w_{K+1}$ und

$$E(a(X_{\tau_{S_K}}) | X_0 = i) + \epsilon \geq w_K(i) + \epsilon \geq v(i) \quad \text{für alle } i \in Z$$

d.h. S_K ist ε -Stoppmenge.

Beweis: Offensichtlich bricht das Verfahren nach höchstens K Schritten ab, d.h. $w_{K+1} = w_K \cdot S_K$ ist eine optimale Stoppmenge zum Stopp-Problem (Z, P, a_K) , weil die S_n nach dem ε -Verfahren der Stoppmengenerweiterung übereinstimmen mit den S_n des Verfahrens der Stoppmengenerweiterung, wenn dieses auf das Problem (Z, P, a_K) angewendet wird.

Da $a_K + \varepsilon \geq a$, folgt dann unmittelbar aus der Tatsache, daß mit w_K auch $w_K + \varepsilon$ ein Supermartingal ist, die Behauptung.

V. Bemerkungen zum Rechenaufwand:

Der Rechenaufwand bei den Verfahren der Stoppmengenänderung besteht wie bei allen Verfahren der Politikiteration hauptsächlich in der Bestimmung der w_n nach (a) bzw. (c) $[(a'), (c'); (a\varepsilon), (c\varepsilon)]$. Zur Berechnung der w_n ($n = 0, 1, \dots$) ist ein lineares Gleichungssystem mit jeweils maximal $z - (n+1)$ Unbekannten zu lösen. Der maximale Rechenaufwand beträgt danach

beim Verfahren der Stoppmengenerweiterung:

(falls $S^* = Z$; und falls bei jedem Schritt die S_n um höchstens einen Zustand erweitert werden)

$$\frac{1}{12} (z^4 + 2z^3 - z^2 + 10z)$$

Rechenoperationen (d.h. Multiplikationen oder Divisionen).

Im Vergleich dazu benötigt die Lösung mit Hilfe des linearen Programms (4), (5), (6) höchstens $z^3 + z^2 - 2$ Rechenoperationen.

Für großen Zustandsraum Z sind also die Verfahren der Stoppmengenänderung dem linearen Programm bezüglich des maximalen Rechenaufwandes unterlegen. Bis $z = 9$ ($z = 4$) ist allerdings das Verfahren der Stoppmengenerweiterung (Stoppmengeneinschränkung) günstiger.

Ist $P > 0$ und $a > 0$ (und hat man keine Vorinformation über die optimale Stoppmenge), dann läßt sich der Mindestrechenaufwand beim linearen Programm in Abhängigkeit von der Zahl r der optimalen Stoppzustände angeben, nämlich

$$z^3 + (2 - r)z^2 + (1 - r)z$$

(man kann sich das leicht anhand des (dualen) Simplextableaus überlegen). - Der maximale Rechenaufwand beim Algorithmus der Stoppmengenerweiterung, falls $|S^*| \leq 3$ ist nun kleiner als

$$z^3 - 3z^2 + 4z$$

d.h. bei endlichem Zustandsraum ist, falls $|S^*| \leq 3$ der maximale Rechenaufwand für das Verfahren der Stoppmengenerweiterung geringer als der minimale Aufwand bei der Lösung mit dem linearen Programm.

VI. Bemerkungen

- (1) Diese Abschätzungen des Rechenaufwands betrachten immer den ungünstigsten Fall für die Verfahren der Stoppmengenerweiterung, daß nämlich die jeweilige Stoppmenge um höchstens einen Zustand abgeändert wird. In den meisten Fällen, insbesondere beim ϵ -Verfahren wird jedoch die Änderung mehrere Zustände betreffen.
- (2) Ein wesentlicher Unterschied zu den anderen Verfahren der Politikiteration besteht darin, daß das Verfahren der Stoppmengenerweiterung eine weitaus bessere Abschätzung der maximalen Schrittzahl (nämlich z) zuläßt als die bisher üblichen Verfahren (vgl. z.B. [2]). Außer in speziellen Problemen (z.B. im absolut monotonen Fall (vgl. [2])) kann man nämlich als obere Grenze der erforderlichen Schrittzahl lediglich 2^N angeben.

- (3) Eine direkte Bestimmung von ϵ -Politiken (d.h. ohne den exakten Spielwert zu kennen) wie sie nach dem ϵ -Verfahren der Stoppmengenerweiterung (a ϵ), (b ϵ), (c ϵ) möglich ist, war dem Verfasser bisher nicht bekannt.
- (4) Das ϵ -Verfahren ist auch bei abzählbarem Zustandsraum anwendbar (vgl. [4]).
- (5) Eine geringe Abänderung ermöglicht die Anwendung des Verfahrens der Stoppmengenerweiterung auch bei Erneuerungsproblemen (vgl. [5]).

VII. Beweis von Satz 2:

Voraussetzung: die Markoff-Kette sei absorbierend. Die Randzustände (d.h. die absorbierenden Zustände) werden mit R bezeichnet.

Lemma: für alle $n = 0, 1, \dots, z$ gilt:

S_n ist Teilmenge der Stützmenge S^*
 w_n ist Supermartingal

- (a) $S_0 \subseteq S^*$: klar (vgl. z.B. [3] Seite 95)
 w_0 Supermartingal:

$$(1) \quad w_0(i) = \sum_{j \in Z} p_0(i, j) w_0(j) \quad (\text{vgl. [2] Seite 24})$$

für $i \in R$ oder $i \notin S_0$ gilt $p_0(i, j) = p(i, j)$ d.h.

$$w_0(i) = \sum p(i, j) w_0(j)$$

für $i \in S_0 \setminus R$:

$$\begin{aligned} w_0(i) &= \max \{a(k); k \in Z\} \\ &\geq \sum p(i, j) \max \{a(k); k \in Z\} \\ &= \sum p(i, j) \max \{w_0(k); k \in Z\} \\ &\geq \sum p(i, j) w_0(j) \end{aligned}$$

d.h. $w_0 \geq Pw_0$ also ist w_0 ein Supermartingal.

- (b) Sei nun w_n ein Supermartingal und $S_n \subseteq S^*$.
 Zu zeigen ist w_{n+1} ist Supermartingal und
 $S_{n+1} \subseteq S^*$.

- (ba) $S_{n+1} \subseteq S^*$:

Es sind nur Zustände $i \in S_{n+1} \setminus S_n$ auf ihre Zugehörigkeit zu S^* zu prüfen.

Man überlegt sich leicht, daß der Spielwert v ein Supermartingal (weil v ein Supermartingal bezüglich der Markoff-Kette mit der Übergangsmatrix P ist) und w_n ein Martingal (vgl. (1)) bezüglich einer Markoff-Kette mit der Übergangsmatrix P_n *) ist.

Aus Theorem 2 von [8] folgt dann, da $v(i) = w_n(i)$ für alle bezüglich der Matrix P_n absorbierenden Punkte, daß

$$v \geq w_n$$

also ist

$$g := v - w_n$$

nicht negativ.

Andererseits ist v ein Martingal bezüglich der Übergangsmatrix $P(S^*)$ während w_n ein Supermartingal ist; d.h. g ist ein Submartingal bzgl. $P(S^*)$. Aus einem Hilfssatz von Bauer ([1] Hilfssatz 2) folgt, daß nicht negative Submartingale ihr Maximum auf den Absorptionszuständen annehmen, d.h. jeder Zustand i für den gilt: $v(i) - w_n(i) = \max \{v(j) - w_n(j); j \in Z\}$ liegt in der Stützmenge S^* . Das war zu zeigen.

(bb) w_{n+1} ist Supermartingal.

für Zustände $i \in S_{n+1}$ gilt $p_{n+1}(i, j) = p(i, j)$ und deshalb auch

$$\begin{aligned} w_{n+1}(i) &= \sum p_{n+1}(i, j) w_{n+1}(j) \\ &= \sum p(i, j) w_{n+1}(j) \end{aligned}$$

für Zustände $i \in S_{n+1}$ gilt nach (ba)

$$w_{n+1}(i) = v(i) = a(i) .$$

) $P_n = P(S_n)$ bzw. $P(S^)$ sind die folgenden stochastischen Matrizen $P(S) = (p(S; i, j))_{i, j \in Z}$ mit

$$p(S; i, j) = \begin{cases} p(i, j) & i \in S \\ 0 & i \in S, i \neq j \\ 1 & i \in S, i = j \end{cases}$$

Daraus folgt wieder mit dem oben zitierten Satz von Kemeny und Snell [8] (weil v ein P_{n+1} Supermartingal ist)

$$v \geq w_{n+1}$$

und daraus

$$w_{n+1}(i) = v(i) \geq \sum p(i,j)v(j) \geq \sum p(i,j)w_{n+1}(j)$$

Das war zu zeigen.

Das Lemma ist somit bewiesen.

Der Beweis des Satzes ergibt sich nun wie folgt:

Die erste Behauptung ist trivial.

Die zweite Behauptung erhält man aus dem Lemma, denn

$$w_n \text{ ist Supermartingal und } w_n \geq a$$

also $w_n \geq v$ (vgl. Satz 7).

Andererseits gilt nach Definition des Spielwertes v

$$(2) \quad v(i) \geq E(a(X_{\tau_{S_n}}) | X_0 = i) = w_n(i)$$

d.h. $w_n = v$, sobald $w_n \geq a$.

Daß S_n eine optimale Stoppmenge ist, ist nach (2) trivial.

Bisher wurde eine absorbierende Markoff-Kette vorausgesetzt. Bauer [1] hat nun gezeigt, daß die Semimartingale einer Markoff-Kette (d.h. Sub- bzw. Supermartingale) die Zustände minimaler Absorptionsmengen (ergodischer Klassen) A_t nicht trennen (d.h. sie sind auf diesen Mengen konstant).

Wenn man dies beachtet, kann man die Sätze von Kemeny und Snell [8] übernehmen, falls man nur jeweils "minimale Absorptionsmenge" für "absorbierender Zustand" (Randzustand) setzt. Die benötigten Sätze von Bauer [1] sind schon für minimale Absorptionsmengen formuliert. Es gilt also das Lemma und damit Satz 9 auch für nicht absorbierende Markoff-Ketten, falls man die minimalen Absorptionsmengen A_t zuvor bestimmt hat und für alle $i \in A_t$

$$a'(i) = \max \{a(j); j \in A_t\}$$

setzt. Das Stopp-Problem (Z, P, a') ist dann zu lösen. Die Ermittlung der A_t kann auch unterbleiben, denn als Anfangsstoppmenge kann statt S_0 auch (vgl. (ac))

$$S'_0 = \{i : a(i) = \max \{a(j); j \in Z\}\}$$

gewählt werden, wie man sich leicht anhand des obigen Beweises überlegt.

Literatur:

- [1] Bauer, H.: Supermartingale und Choquet-Rand.
Arch. Math. 12, 210 - 223; 1961
- [2] Derman, C.: Finite State Markovian Decision Processes.
New York, London; 1970
- [3] Dynkin, E.B., Juschkevitsch, A.A.: Sätze und Aufgaben
über Markoffsche Prozesse. Berlin, Heidelberg,
New York; 1969
- [4] Emrich, O.: Algorithmen zur Bestimmung optimaler
Stoppregeln bei endlichen Markoff-Ketten.
Dissertation Karlsruhe, 1971.
- [5] Emrich, O.: Bestimmung optimaler Stoppmengen bei
binären Markoffschen Erneuerungsprozessen mit Hilfe von Verfahren der Politik-
iteration.
Erscheint in Operations Research Verfahren
XII, 1972
- [6] Emrich, O., Göppl, H.: Zur Lösung einiger Investitionsprobleme mittels Markoffscher Entscheidungsprozesse.
Erscheint in Operations Research Verfahren
Band XI, 1972
- [7] Henke, M.: Sequentielle Auswahlprobleme bei Unsicherheit. Ökonometrische Studien, Bd. 2,
Meisenheim; 1970.

Prognose- und Schätzverfahren

Die Bedarfsanalyse als Hilfsmittel für technisch-wirtschaftliche Planungen

von H. M. Dathe, Ottobrunn

Zusammenfassung

Zur Klärung der Bewertung und der Nutzung industrieller Produkte ist man häufig auf subjektive Beurteilungen angewiesen. Dies gilt auch für das "Modal Split"-Problem, d.h. für die Frage, nach welchen Gesichtspunkten ein Benutzer unter mehreren konkurrierenden Geräten für den gleichen Verwendungszweck (z.B. Verkehrs- oder Kommunikationsmittel) seine Wahl trifft. Es kommt dabei darauf an, eine Befragung von Benutzern in der Weise vorzunehmen, daß man die erhaltenen subjektiven Urteile als Grundlage für ein Bedarfsmodell verwenden kann. Als geeignet hierfür erweist sich die Definition von Bedarfsfaktoren sowie die Ermittlung ihrer Gewichtung und ihrer Erfüllungsgrade bei einzelnen Produktarten. Eine wesentliche Stütze des Bedarfsmodells bilden die gleichzeitig erfaßten objektiven Nutzungsanteile der Produkte.

In einem für Kommunikationsmittel durchgeführten Experiment wurde die Brauchbarkeit dieser Methode nachgewiesen. Die Auswertung - die auch die Streuung der Befragungsergebnisse berücksichtigt - bestätigt u.a. den angenommenen nichtlinearen Zusammenhang zwischen Wertschätzung und Verwendung eines Kommunikationsmittels. Mit Hilfe eines solchen Bedarfsmodells lassen sich Marktprognosen durchführen und vor allem die Auswirkungen von technischen und anwendungsmäßigen Verbesserungen abschätzen.

Summary

In the evaluation of industrial products and their practical applications, subjective judgment is often the only available source of investigation. This is also true when seeking a solution for the "modal split" problem, i.e., when trying to find the reasons that guide a user in his choice among transportation systems, communication systems or other devices. It is important in this context to organize the inquiries to users in such a manner that the design of a demand model can be based on the subjective answers obtained. This is done by defining suitable demand factors and by determining their relative weights and acceptance indices for various kinds of products. To record the utilization shares of these products is essential for cross-checking the demand model.

The usefulness of this method was tested by an experiment carried out for communication devices. The assessment of results - which includes a variance analysis of statistical data - confirms the assumption that the application shares are a nonlinear function of product appraisal. By means of this demand model, market forecasts can be carried out and the consequences of improvements in technology or systems application can be estimated.

1. Einleitung

Im Bereich des Operations Research hat in den letzten Jahren die Auswertung subjektiver Beurteilungen in zunehmendem Maße Beachtung gefunden. Dabei galten die Bemühungen vor allem der Erarbeitung einer stichhaltigen methodischen Fundierung von Verfahren der Punktbeurteilung, der Entscheidungsmatrix, des Relevanzbaumes und der Marktbefragung, vgl. [1]. Subjektive Beurteilungen sind bei korrekter Handhabung im Instrumentarium des Analytikers eine nützliche Ergänzung zu rein quantitativen Methoden, sie sind in der Regel auch für Nichtanalytiker gut überschaubar und verwendbar und sie erlauben am ehesten eine Einbeziehung wesentlicher soziologischer und psychologischer Aspekte in die Betrachtung.

Mit der Auswertung von Benutzer-Befragungen als einer speziellen Klasse subjektiver Beurteilungen befassen wir uns im Folgenden. Im Mittelpunkt der Untersuchung steht dabei das "Modal Split"-Problem, das vor allem im Zusammenhang mit Nahverkehrs- und Fernverkehrsfragen eine große aktuelle Bedeutung besitzt: Nach welchen Kriterien wählt ein Verkehrsteilnehmer die für ihn günstigste Beförderungsart aus, d.h. welche Faktoren sind für die relative Benutzungshäufigkeit der einzelnen Verkehrsmittel und deren Änderung maßgebend? Unter welchen Bedingungen kann die Bahn auf Mittelstrecken Reisende zurückgewinnen, und besteht Aussicht, das Großstadtpublikum zum stärkeren Verzicht auf die Benutzung des eigenen Pkw zu veranlassen? Ganz analoge Fragestellung treten beim Gütertransport, bei der Nachrichtenübertragung (Kommunikationsmittel) und bei zahlreichen anderen Geräteklassen und konkurrierenden Wirtschaftsgütern auf.

Für den Versuch einer möglichst wirklichkeitsnahen Untersuchung eines solchen "Modal Split"-Problems mit Hilfe einer Befragung sollte man zumindest von folgenden Voraussetzungen ausgehen :

- begrenztes, klar umrissenes Aufgabengebiet der konkurrierenden Mittel oder Geräte,
- Befragung einer ausreichenden Zahl kompetenter Benutzer,
- zusätzliche Ermittlung der Nutzungsanteile durch die Befragung, um die Ergebnisse der Auswertung überprüfen zu können.

Auf dieser Grundlage kann eine sog. Bedarfsanalyse durchgeführt werden, deren grundsätzlicher Ablauf in Bild 1 skizziert ist.

Für das betrachtete Aufgabengebiet wird eine größere Anzahl (8 oder 10) Bedarfsfaktoren definiert, durch die alle vom Benutzer an ein Kommunikationsmittel (oder ein Verkehrsmittel) gestellten Anforderungen charakterisiert werden. Der Befragte wird gebeten, die relative Bedeutung der einzelnen Faktoren durch Angabe einer Gewichtung oder einer Rangfolge auszudrücken. Außerdem schätzt er Erfüllungsgrade, d.h. die Prozentsätze ab, mit denen einzelne Mittel die durch die Bedarfsfaktoren ausgedrückten Wünsche erfüllen. Die "Mittel" sind dabei 3 bis 5 gegenwärtige Konfigurationen gebräuchlicher Kommunikations- oder Verkehrsmittel.

Die dritte Angabe des Befragten bezieht sich auf die Nutzungsanteile, die nach seinen Erfahrungen auf die 3 bis 5 Mittel entfallen. Aufgabe des Analytikers ist es nun, ein Bedarfsmodell zu entwickeln, das die ermittelten Nutzungsanteile auf Grund der Gewichtungen und Erfüllungsgrade der Bedarfsfaktoren erklärt. Bei dieser Strukturanalyse des Modells sind additive und multiplikative, lineare und nichtlineare Verknüpfungen in Betracht zu ziehen, soweit sie logisch begründet sind. Da die Befragungsergebnisse

statistische Unterlagen darstellen, sind neben den Erwartungswerten noch die Streuungen durch eine Varianzanalyse auszuwerten.

Die Struktur des Bedarfsmodells (oder ggfs. die Definition der Bedarfsfaktoren) ist korrekturbedürftig, solange noch keine ausreichende Übereinstimmung zwischen dem theoretischen Verwendungspotential und den praktischen Nutzungsanteilen erzielt ist. Wird das Bedarfsmodell jedoch durch diesen Prozeß verifiziert, so spricht dies für eine Verwendung als Hilfsmittel für weitere Planungen. Man kann z. B. abschätzen, wie sich zukünftige Änderungen an einzelnen Verkehrsmitteln auf ihre Erfüllungsgrade auswirken werden, und die geänderten Verwendungspotentiale der einzelnen Mittel hieraus mit Hilfe des Bedarfsmodells berechnen. Befragungen der heutigen Benutzer über zukünftige Verwendungsanteile sind erfahrungsgemäß wenig zuverlässig. Vermag man jedoch bei Anwendung der beschriebenen Methode den Einfluß organisatorischer oder technischer Änderungen auf die Erfüllungsgrade realistisch abzuschätzen, so besteht Grund zu der Annahme, daß man damit zu korrekteren Prognoseergebnissen gelangt.

Natürlich sind auch Untersuchungen zu einer umgekehrten Fragestellung möglich: Wie stark müssen wichtige Benutzungseigenschaften von Geräten verbessert werden, damit man größere Verwendungsanteile (z.B. bei der Bahn) erzielt?

Das Prinzip der Lösung des "Modal Split"-Problems unter Anwendung einer Bedarfsanalyse geht auf A. Sommers und F. Leimkuhler [2] zurück. Das hiernach in [1] beschriebene Verfahren wird im folgenden auf Grund von experimentellen Erfahrungen erweitert und modifiziert. Bei den Darlegungen in den nächsten Abschnitten erweist sich die Bezugnahme auf ein konkretes Beispiel aus Illustrationsgründen als zweckmäßig.

2. Die Definition von Bedarfsfaktoren

Ausgangspunkt der Betrachtung muß ein begrenztes, klar umrissenes Aufgabengebiet sein, in dem mehrere Hilfsmittel oder Geräte in Ergänzung oder Konkurrenz zueinander eingesetzt werden. Als Beispiel wählen wir die aktive, dienstliche Benutzung von Kommunikationsmitteln auf Entfernungen zwischen etwa 100 und 1000 km. Mit der Einschränkung "aktiv" ist die Voraussetzung verbunden, daß der Benutzer die Art des Kommunikationsmittels frei wählt, daß also Antworten und Reaktionen auf fremde Anfragen nicht berücksichtigt werden. Als Kommunikationsmittel kommen in Betracht:

- a. Telefon, d.h. unmittelbar zugängliche Geräte, wobei die Vermittlung von Ferngesprächen durch eine Zentrale erfolgt,
- b. Fernschreiber, d.h. Geräte, die von einer Zentrale auf Grund von schriftlichen Vorlagen bedient werden,
- c. Briefe, die im Schreibbüro der Abteilung oder Gruppe angefertigt werden,
- d. Dienstreisen, die zur Durchführung von wichtigen, ausgedehnten oder mehrseitigen Kommunikationen auf Grund eines Reiseantrags durchgeführt werden.

Die einzelnen Aspekte des gesamten Kommunikationsbedarfs sollen durch die folgenden Bedarfsfaktoren F_j ($j=1,2,\dots,n$) zum Ausdruck kommen:

- F_1 Verständigung:
- Der Gesprächspartner soll Wortlaut, Sinn und Hintergrund der Nachricht bzw. Anfrage gut verstehen.

- F₂ Zusammenarbeit :**
Die Bereitwilligkeit des Gesprächspartners zur Zusammenarbeit soll erhalten, geweckt und u.U. verstärkt werden.
- F₃ Nachdruck :**
Der Gesprächspartner soll mit Nachdruck zur Antwort bzw. Aktion veranlaßt werden.
- F₄ Wiederholungen :**
Das Mittel muß in schwierigen Kommunikationsfällen für eine Wiederholung des Nachrichtenaustausches psychologisch geeignet und dabei erfolgversprechend sein.
- F₅ Belästigung :**
Die Belästigung in qualitativer wie quantitativer Hinsicht (Zeitbedarf) soll für den angesprochenen Partner so klein wie möglich sein.
- F₆ Effizienz :**
Der Zeit- und Kostenaufwand für den Benutzer soll in einem vernünftigen Verhältnis zum Nutzen bzw. zur Bedeutung der Kommunikation stehen.
- F₇ Verzögerungen :**
Die Kommunikation soll mit relativ geringen Verzögerungen im Informationszyklus zustande kommen, möglichst geringe Kollisionen mit dem Terminkalender verursachen und den Partner zu einer raschen Reaktion veranlassen.
- F₈ Begrenzungen :**
Die kostenmäßig und durch andere Termine gesetzten Begrenzungen des Umfangs einer Nachricht sollen den Erfolg einer

Kommunikation möglichst wenig beeinträchtigen.

F₉ Bequemlichkeit:

Sieht man von Zeit- und Kostengesichtspunkten ganz ab, so soll der Kommunikationsvorgang angesichts anderer wichtiger Aufgaben keine besonderen Umstände erfordern, d.h. für den Benutzer bequem verlaufen.

Das Konzept der Bedarfsfaktoren ist nicht frei von Mängeln: der Bedeutungsinhalt einiger Faktoren läßt sich nur schwer vom jeweiligen Mittel trennen, ihr Bedeutungsumfang überschneidet sich teilweise mit demjenigen anderer Faktoren. Hierbei handelt es sich jedoch um übliche Modellvereinfachungen, die angesichts anderer Unsicherheiten in Kauf genommen werden können und die den Wert der Analyse nicht entscheidend zu beeinträchtigen brauchen.

3. Gewichtungen und Erfüllungsgrade

Bei der Bewertung der Bedarfsfaktoren unterscheidet man wie bei einer Punktbeurteilung grundsätzlich zwei verschiedene Parameter:

die Gewichtungen p_j ($j=1,2,\dots,n$) der einzelnen Bedarfsfaktoren j , die von den Produkten unabhängig sind, und

die Erfüllungsgrade a_{ij} ($i=1,2,\dots,m$), welche angeben, inwieweit ein Produkt i (eines der Kommunikationsmittel) einen Bedarfsfaktor j erfüllt.

Alle Unterschiede in der Bedeutung der Bedarfsfaktoren für einzelne Produkte sollen mit in den a_{ij} zum Ausdruck kommen. Die Gewichtungen p_j legen

die Wertigkeit der Faktoren innerhalb des betrachteten Kommunikationsbedarfs fest, sie sind völlig losgelöst von der Bedarfsdeckung durch die Mittel zu bestimmen.

Die Größen p_j und a_{ij} sind die Parameter des Bedarfsmodells, das zur Beschreibung des Verwendungspotentials v_i jedes Produktes i dient:

$$v_i = f(p_j, a_{ij}) \quad (1)$$

Unser Ziel ist die Bestimmung der Funktion f für Kommunikationsmittel mit Hilfe einer Bedarfsanalyse, die auf Befragungsdaten für die Gewichtungen, Erfüllungsgrade und Nutzungsanteile beruht. Befragt wurden 20 erfahrene Benutzer von Kommunikationsmitteln, die alle der gleichen Firma angehören. Diese Befragung diente in erster Linie der Erprobung des Verfahrens, weniger einer gezielten Bedarfsforschung.

Für die Festlegung der Gewichtung der 9 Bedarfsfaktoren wurden die Befragten gebeten, eine Summe von 90 Punkten zu vergeben, wobei der bedeutendste Faktor durch die relativ größte Punktzahl zu kennzeichnen war. Bei 90 Punkten kam im vorliegenden Fall sehr einfach eine lineare Rangeinstufung

$$18 - 16 - 14 - 12 - 10 - 8 - 6 - 4 - 2$$

eine Drei-Klassen-Einteilung

$$15 - 15 - 15 - 10 - 10 - 10 - 5 - 5 - 5$$

oder jede andere Aufteilung in Einzelpunkte vorgenommen werden.

Das Befragungsergebnis für die Gewichtungen ist in Tabelle 1 aufgelistet. Außer der mittleren Punktzahl für jeden Bedarfsfaktor ist jeweils die kleinste und die größte vergebene Punktzahl angegeben. Diese drei Werte weisen

eine bemerkenswerte Konsistenz auf, was dem bereits wirksamen Gesetz der großen Zahlen ebenso wie dem Sachverstand der Befragten zu danken ist. Praktisch läßt sich keine Unterscheidung zwischen Faktoren mit kleiner und mit großer Streuung der Gewichtung treffen. Wir wollen es deshalb als zulässig ansehen, mit den mittleren Punktzahlen allein als typischem Befragungsergebnis weiterzuarbeiten. Die so ermittelte Gewichtung ist wesentlich besser fundiert, als sie es auf Grund einer Abschätzung durch nur zwei bis drei Benutzer wäre. Aus Gründen der Zweckmäßigkeit wird die auf die Summe 90 ausgerichtete mittlere Punktzahl so auf die Gewichtungen p_j umgerechnet, das deren Summe 100 beträgt. Den Bedarfsfaktoren Verständigung und Zusammenarbeit wird eine doppelt so große Bedeutung wie vielen anderen Faktoren beigemessen.

Schwieriger als die Bestimmung der Gewichtung ist die Ermittlung der Erfüllungsgrade, da man hierbei mit weit stärker voneinander abweichenden Meinungen der Befragten rechnen muß. Zur Erleichterung der Einstufung wird das Intervall 0 bis 1 wie folgt in Abschnitte mit zugehörigen verbalen Begriffen unterteilt:

sehr schlecht	0 bis 0,2
schlecht	0,2 bis 0,4
mittel	0,4 bis 0,6
gut	0,6 bis 0,8
sehr gut	0,8 bis 1,0

Die Befragten werden aufgefordert, jeweils eines der so entstehenden Felder anzukreuzen. Kommt ein Befragter auf Grund verschiedener Erfahrungsfälle mit einem Kommunikationsmittel zu unterschiedlichen Beurteilungen, so trägt er nur den Mittelwert dieser Urteile ein.

Die Häufigkeiten der Antworten von 20 Befragten sind in Tabelle 2 einge-

tragen. Um Mißverständnisse bei den Befragten weitgehend zu vermeiden, wurden die schraffierten Bewertungsfelder vor der Befragung ausgespart, weil sie bei rationaler Betrachtung für eine Antwort sicher nicht in Frage kommen. Die Meinungsvielfalt der 20 Befragten wird aus der Tatsache ersichtlich, daß sie ihre Antworten auf 87 % der zur Verfügung stehenden Felder verteilten. Wenn auch die Streuungen groß sind, so zeichnen sich dennoch gewisse Beurteilungsschwerpunkte ab.

Die Auswertung geht deshalb auch von dem Gedanken aus, daß die maßgebende Beurteilung eher Ausdruck einer Mehrheitsmeinung als der mit ungerechtfertigter Genauigkeit zu errechnende Mittelwert aus allen Antworten sein sollte. Das Bild aller Antworten, Tabelle 2, ist als Hintergrundinformation wertvoll, für Zwecke des Bedarfsmodells wird jedoch eine stärker zusammengefaßte Aussage über die Erfüllungsgrade benötigt. Sie ist in Bild 2 gegeben, wo nur noch zwischen Mehrheitsbeurteilungen und Feldern unterschieden wird, die der Beurteilung stärkerer Minderheiten entsprechen.

Für den Übergang von Tabelle 2 auf Bild 2 wurden folgende Kriterien angewandt, die empirischen Charakter besitzen:

- a) Ein Erfüllungsgrad a_{ij} erhält ausschließlich Mehrheitsfelder, wenn sich auf ein Feld mindestens 45 % bzw. auf 2 bis 3 nebeneinanderliegende Felder mindestens 60 % der Antworten konzentrieren. Die nebeneinanderliegenden Mehrheitsfelder dürfen sich außerdem in ihrem Wert nicht um mehr als 5 % der Antworten unterscheiden.
- b) Ein Erfüllungsgrad a_{ij} erhält sog. Minderheitsfelder neben Mehrheitsfeldern, wenn sich auf ein Minderheitsfeld mindestens 25 % bzw. auf zwei Felder mindestens 50 % der Antworten konzentrieren. Die Minderheitsfelder müssen sich in ihrem Wert von den danebenliegenden Mehrheitsfeldern um mehr als 5 % unterscheiden.

Bild 2 vermittelt einen besseren Überblick über die erhaltenen Erfüllungsgrade und ihre Varianzbreiten. Natürlich könnte man noch unterscheiden, ob ein einzelnes Mehrheitsfeld von 95 % oder z.B. nur von 60 % der Antworten getragen wird, doch wurde hierauf in der vorliegenden Analyse verzichtet.

4. Ermittlung der Nutzungsanteile

Die auf die betrachteten vier Kommunikationsmittel entfallenden relativen Nutzungsanteile sind neben den Gewichtungen und den Erfüllungsgraden der Bedarfsfaktoren die dritte Größe, die durch die Befragung ermittelt wird. Es kommt bei diesen prozentualen Nutzungszahlen nicht auf Unterschiede hinsichtlich Qualität, Zeit- und Kostenaufwand der Kommunikationen, sondern lediglich auf Häufigkeitszahlen an. Dabei war davon auszugehen, daß auf einer Dienstreise - im Gegensatz zu den anderen Mitteln - zwei oder mehr aktive Kommunikationen erfolgen können, wodurch sich die für Dienstreisen einzusetzende relative Häufigkeitszahl entsprechend erhöht.

Die Antworten von 19 Befragten sind in Tabelle 3 angegeben. Durch Summierung und Mittelwertbildung wurden daraus die Nutzungsanteile errechnet, die somit für das Kollektiv der Befragten gelten. Zweifellos haben einzelne Teilnehmer an der Befragung eine größere Erfahrung mit Kommunikationen als andere, die pro Monat eine geringere Zahl von Telefongesprächen, Briefen usw. zu bewältigen haben. Da diese Unterschiede im vorliegenden Fall jedoch nicht sehr ausgeprägt waren und außerdem nur schwer korrekt zu erfassen sind, fanden sie keine Berücksichtigung.

Wenden wir uns nun dem Aspekt der Streuungen zu, so ist festzustellen, daß die Verteilung der Einzelantworten um die berechneten Mittelwerte wenig aussagt, daß hingegen die Genauigkeit, mit der jeder Befragte seine Angaben gemacht hat, nicht unwesentlich ist. Aus den von jedem Befragten angegebenen 4 Zahlen ist unschwer zu erkennen, ob der Betreffende auf Zehntel, Zweier, Fünfer oder Zehner aufgerundet hat. Daraus ergibt sich das Unsicherheitsintervall $(c-b)$ in der vorletzten Spalte von Tabelle 3.

Für die Verteilung der Häufigkeit, die von einem Befragten für ein Kommunikationsmittel angegeben wird, ist somit von einer Rechteckverteilung mit der Breite $(c-b)$ auszugehen, für deren Varianz gilt:

$$\sigma^2 = \frac{(c-b)^2}{12} \quad (2)$$

Betrachtet man die Verteilung zweier unabhängiger Zufallsgrößen x und y , deren Erwartungswerte zu summieren sind:

$$E(z) = E(x) + E(y) , \quad (3)$$

so erhält man auch die Varianz der resultierenden Verteilung z durch Addition der Einzelvarianzen:

$$\sigma^2(z) = \sigma^2(x) + \sigma^2(y) \quad (4)$$

Nach dem Zentralen Grenzwertsatz nähert sich die durch eine solche Summierung erhaltene resultierende Verteilung einer Normalverteilung, wenn die Zahl n der Einzelverteilungen genügend groß ist. Im vorliegenden Fall ($n=19$) sind die genannten Voraussetzungen weitgehend erfüllt. In der letzten Spalte von Tabelle 3 wird die Varianz - die für alle vier Nutzungsanteile gleich ist - zu $\sigma_u^2 = 84$ berechnet. Die so ermittelten

Normalverteilungen für die Nutzung der Kommunikationsmittel sind in Bild 6 eingetragen, das später erläutert wird.

5. Entwicklung des Bedarfsmodells

Die Nutzungsanteile dienen dem Zweck, das zu entwickelnde Bedarfsmodell auf seine Wirklichkeitstreue zu prüfen. Der Struktur dieses Modells - die nach zweimaligem Durchlaufen der in Bild 1 dargestellten Schleife gefunden wurde - liegen folgende Gedanken zugrunde:

- a. Das Modell berücksichtigt nur die Erwartungswerte von Gewichtungen und Erfüllungsgraden, während die Streuungen gesondert in einer Varianzanalyse untersucht werden.
- b. Für jeden Bedarfsfaktor werden die mittleren Erfüllungsgrade a_{ij} der vier Kommunikationsmittel miteinander verglichen.
Das Minimum unter ihnen dient als Bezugsgröße:

$$a_{0j} = \min_i \{ a_{ij} \} \quad (5)$$

Die Differenz α_{ij} der übrigen Erfüllungsgrade von dieser Bezugsgröße

$$\alpha_{ij} = a_{ij} - a_{0j} \quad (6)$$

wird in der weiteren Rechnung verwendet. Durch diese Differenzbildung für jeden Bedarfsfaktor wird sichergestellt, daß die Überlegenheit eines Mittels im Vergleich zu anderen in sinnvoller Weise berücksichtigt wird.

- c. Die erhaltenen Größen α_{ij} werden mit der jeweiligen Gewichtung p des Bedarfsfaktors multipliziert, bevor eine Summierung dieser Produkte durchgeführt wird. Eine Ausnahme bildet lediglich der Bedarfsfaktor Nr. 6 "Effizienz", der zur erwähnten Summe nicht addiert wird, sondern mit ihr multipliziert wird. Dies entspricht dem Charakter der Effizienz als dem Verhältnis von Nutzen zu Aufwand eines Mittels; sie kann nicht durch andere Vorteile kompensiert werden, sondern vergrößert oder verkleinert das Verwendungspotential eines Mittels unmittelbar. Eine Differenzbildung unterbleibt bei der Effizienz, der Erfüllungsgrad α_{i6} ist vielmehr unmittelbar als Faktor zu verwenden. Die Gewichtung p_6 kürzt sich aus der Formel heraus (vgl. Bild 5).
- d. Die durch die beschriebenen Rechenoperationen erhaltenen Größen sind noch nicht notwendigerweise gleich dem echten Verwendungspotential der einzelnen Kommunikationsmittel. Wir wollen sie als Wertschätzung w_i bezeichnen und sie so normieren, daß ihre Summe 100 % ergibt.

Für die Wertschätzung w_i gilt somit folgende Beziehung:

$$w_i = \alpha_{ie} \sum_j p_j \cdot \alpha_{ij} \quad (7)$$

Dabei ist α_{ij} nach Gl. (6) einzusetzen, der Erfüllungsgrad α_{ie} für die Effizienz ist identisch mit α_{i6} . Die Summierung über alle Bedarfsfaktoren ist für $j = 1$ bis 9 mit Ausnahme von Nr. 6 durchzuführen. Die nach Umrechnung von w_i auf Prozentsätze erhaltene Größe bezeichnen wir mit \bar{w}_i , vgl. Bild 5.

Die Rechenschritte, die zur Ermittlung der Erwartungswerte \bar{w}_i der Wertschätzung führen, sind für das Zahlenmaterial unseres Beispiels in Bild 3

tabellarisch zusammengestellt. Dadurch, daß bei den α_{ij}^* nur die Differenz der Erfüllungsgrade berücksichtigt wird, geht nur die relative Überlegenheit eines Mittels bezüglich eines Bedarfsfaktors in die Rechnung ein. Hätte ein Mittel bei keinem der Bedarfsfaktoren einen Vorzug gegenüber einem anderen Mittel aufzuweisen, so wären alle seine $\alpha_{ij}^* = 0$, d.h. ihm käme kein Verwendungspotential zu - wie dies auch der Wirklichkeit entspricht. Unter den vier betrachteten Kommunikationsmitteln ergibt sich für das Telefon mit $\bar{w}_T = 38,1$ % die größte und für den Fernschreiber (Telex) mit 12,2 % die geringste Wertschätzung.

Befassen wir uns nun mit dem Zusammenhang zwischen der Wertschätzung \bar{w}_i und dem Verwendungsanteil v_i eines Produktes, vgl. Bild 4. Die praktische Erfahrung lehrt, daß diese beiden Größen nicht direkt proportional zueinander sind, wie es der gestrichelt gezeichneten Diagonale im Diagramm Bild 4 entspräche. Ändert sich die Wertschätzung für ein Produkt von 40 auf 60 %, so steigt auch das Verwendungspotential rasch über die 50 %-Grenze hinweg. Ob dagegen die Wertschätzung bei 85 % oder bei 95 % liegt, ändert für viele Benutzer nichts an der Tatsache, daß es sich um ein "praktisch allein verwendbares" Produkt handelt. Ganz Entsprechendes gilt für das untere Ende der Skala, wo eine Erhöhung der Wertschätzung von 5 auf 15 % für die meisten Benutzer noch nicht ausschlaggebend ist.

Wenn die Tendenz dieses Zusammenhangs auch ohne weiteres einleuchtet, so ist doch noch nichts über seine Stärke bekannt. Es erscheint sinnvoll, zur Beschreibung des Zusammenhangs eine möglichst einfache, stetige Funktion zu verwenden, die aus der Psychologie unter dem Stichwort "Lernkurve" [4] bekannt ist. Die erste Ableitung einer solchen Funktion folgt dem Ansatz

$$\frac{d v_i}{d \bar{w}_i} = k \cdot \bar{w}_i \cdot (1 - \bar{w}_i) \quad (8)$$

Integriert man diesen Ausdruck und beachtet man, daß für $\overline{w}_i = 1$ Verwendungsanteil $v_i = 1$ erreicht sein soll, so erhält man $k = 6$. Hierfür ist die symmetrische Kurve nach Gl. (8) in Bild 4, links oben, aufgetragen. Die Abhängigkeit zwischen Verwendung v_i und Wertschätzung \overline{w}_i folgt der Funktion

$$v_i = 3 \overline{w}_i^2 - 2 \overline{w}_i^3, \quad (9)$$

die im großen Diagramm in Bild 4 als S-Kurve dargestellt ist. Wir wollen diese Kurve, die das Nachhinken der Verwendung bei geringen und ihr Vorseilen bei hohen Wertschätzungsanteilen wiedergibt, als Präferenzkurve bezeichnen. Die Verwendung einer von Gl. (9) abweichenden (komplizierteren) Funktion würde die Modellergebnisse nur relativ wenig verändern.

Damit haben wir die Aufgabe, die in Gl. (1) beschriebene Funktion f für Kommunikationsmittel unter den hier vorliegenden Voraussetzungen zu bestimmen, erfüllt. Die Gleichungen, die das Bedarfsmodell bilden, sind in Bild 5 zusammengestellt. Wie bei der Wertschätzung \overline{w}_i so wird auch beim Verwendungspotential \overline{v}_i durch Division durch die Summe aller Anteile eine Normierung auf Prozente vorgenommen.

6. Varianzanalyse

Die Bedarfsanalyse wäre unvollständig, wenn sie sich nur auf die Erwartungswerte der ermittelten Größe bezöge und nicht auch deren Streuungen berücksichtigen würde. Die Streuung der Nutzungsanteile wurde in Tabelle 3 ermittelt, vgl. Abschnitt 4. Um eine entsprechende Analyse für die Wertschätzungen durchführen zu können, müssen wir uns zunächst kurz mit der Multiplikation zweier Zufallsgrößen befassen.

Aus den Zufallsgrößen x und y , die unabhängig voneinander sind und beliebige Verteilungen besitzen, werde eine dritte Größe z durch Multiplikation

$$z = x \cdot y \quad (10)$$

gebildet. Bezeichnen wir die Erwartungswerte der Verteilungen von x und y mit $E(x)$ und $E(y)$, so gilt

$$E(z) = E(x) \cdot E(y) \quad (11)$$

Außer den Erwartungswerten seien die Streuungen der beiden Verteilungen bekannt, wir wollen sie mit σ_x^2 und σ_y^2 bezeichnen. Nach der Verschiebungsregel für die Streuung einer Zufallsgröße kann man schreiben:

$$\sigma_x^2 = E \left[x - E(x) \right]^2 = E(x^2) - E^2(x) \quad (12)$$

Um die unbekannte Streuung σ_z^2 der resultierenden Verteilung zu berechnen, entwickeln wir:

$$\begin{aligned} \sigma_z^2 &= E(z^2) - E^2(z) \\ &= E(x^2 \cdot y^2) - E^2(x) \cdot E^2(y) \\ &= E(x^2) \cdot E(y^2) - E^2(x) \cdot E^2(y) \\ &= [\sigma_x^2 + E^2(x)] \cdot [\sigma_y^2 + E^2(y)] - E^2(x) \cdot E^2(y) \end{aligned}$$

und erhalten schließlich:

$$\sigma_z^2 = \sigma_x^2 \cdot \sigma_y^2 + E^2(x) \cdot \sigma_y^2 + E^2(y) \cdot \sigma_x^2 \quad (13)$$

Für die Wertschätzung w_i nach Gl. (7) können wir mit der Abkürzung

$$S_i = \sum_j p_j \alpha_{ij} \quad (14)$$

schreiben :

$$w_i = a_{ie} \cdot S_i, \quad (15)$$

wobei es sich bei a_{ie} und S_i - und damit auch bei w_i - wohlgemerkt bereits um Erwartungswerte handelt. Wenden wir nun Gl. (13) mit den bei ihrer Ableitung getroffenen Voraussetzungen an, so haben wir für die Varianz der Wertschätzung

$$\sigma^2(w_i) = \sigma^2(a_{ie}) \cdot \sigma^2(S_i) + a_{ie}^2 \cdot \sigma^2(S_i) + S_i^2 \cdot \sigma^2(a_{ie}) \quad (16)$$

Wie in Abschnitt 3 ausgeführt, betrachten wir die Gewichtungen p_j als Kollektivaussage aller Befragten, d.h. es sind bei p_j keine Varianzen zu berücksichtigen. Mithin ist nach Gl. (14)

$$\sigma^2(S_i) = \sum_j p_j \cdot \sigma^2(\alpha_{ij}), \quad (17)$$

wobei die Streuungen der α_{ij} durch p_j lediglich gewichtet werden.

Die Bestimmung der Einzelstreuungen der α_{ij} wird durch die Tatsache erschwert, daß α_{ij} die Differenz der beiden Zufallsgrößen a_{ij} und a_{oj} darstellt. Wir betrachten hier nur die in Bild 2 durch Kreuzschraffur gekennzeichneten Mehrheits-Streubereiche und gehen von der Varianz $\sigma^2(\alpha_{ij}) = 0$ für $\alpha_{ij} = 0$ aus. Bezeichnen wir mit b und c die Eckwerte der zu berücksichtigenden Rechteckverteilung für α_{ij} , so erhalten wir die in Tabelle 4 zusammengestellten Werte für die Varianzbreite $(c - b) \alpha_{ij}$ und für die zugehörigen Streuungen

$$\sigma^2(\alpha_{ij}) = \frac{(c-b)^2 \alpha_{ij}}{12} \quad (18)$$

Die mit diesen Grundlagen nach Gl. (16) berechneten $\sigma^2(w_i)$ sind in der untersten Zeile in Tabelle 4 angegeben. Das letzte Glied in Gl. (16) erweist sich als dominant gegenüber den übrigen Größen, so daß näherungsweise gilt:

$$\sigma^2(w_i) \approx S_i^2 \cdot \sigma^2(a_{ie}) \quad (19)$$

Daraus ergeben sich zwei Folgerungen:

- Änderungen in den Breiten der nach Bild 2 berücksichtigten Streubereiche wirken sich auf $\sigma^2(w_i)$ praktisch nicht aus.
- Eine Rechteckverteilung erweist sich als viel geeigneteres Modell für die Varianz der Wertschätzung als etwa eine Normalverteilung oder eine andere Verteilungsform.

Legt man Rechteckverteilungen für die vier Kommunikationsmittel zugrunde, so bleibt aus Gründen der geometrischen Ähnlichkeit beim Übergang von w_i auf \bar{w}_i , v_i und \bar{v}_i für ein Mittel das Verhältnis des σ -Wertes zum Erwartungswert konstant. Danach kann man die $\sigma(w_i)$ nach Tabelle 4 in $\sigma(\bar{v}_i)$ -Werte umrechnen und man erhält:

Telefon	$\sigma(\bar{v}_T) = 6,95$
Telex	$\sigma(\bar{v}_X) = 1,19$
Brief	$\sigma(\bar{v}_B) = 2,07$
Dienstreise	$\sigma(\bar{v}_D) = 2,49$

Der größte σ -Wert tritt beim Telefon auf, da hier ein großer Erwartungswert der Wertschätzung mit einer relativ großen a_{ie} -Streuung zusammentrifft, vgl. Gl. (19).

Für die zugehörigen Varianzbreiten der Rechteckverteilungen gilt in Analogie zu Gl. (18)

$$(c-b)_{\bar{v}_i} = \sqrt{12} \cdot \sigma(\bar{v}_i)$$

7. Vergleich von Verwendungspotential und Nutzung

Zur Überprüfung des entwickelten Bedarfsmodells können wir nun die zuletzt berechneten Verwendungspotentiale der vier Kommunikationsmittel mit den Nutzungsanteilen lt. Befragung, vgl. Abschnitt 4, vergleichen. In Bild 6 sind die Häufigkeitsverteilungen über der Achse der Prozentanteile aufgetragen. Die Normalverteilungen für die wirkliche Nutzung der vier Mittel haben die gleiche Gestalt, da ihnen die gleiche Streuung zugrunde liegt. Die Höhen der Rechteckverteilungen für die Verwendungspotentiale sind so zu wählen, daß der Flächeninhalt mit dem der Normalverteilungen identisch ist.

Wie diese Auftragung anschaulich zeigt, ist die Übereinstimmung in allen vier Fällen gut, wodurch das Bedarfsmodell in seiner endgültigen Form, Abschnitt 5, bestätigt wird. Aus der Tatsache, daß das Verwendungspotential über oder unter dem Nutzungsanteil liegt, kann man kaum tiefgreifende Folgerungen ziehen. Wegen der Betrachtung von Prozentsätzen bei den einzelnen Mitteln muß die Summe aller Abweichungen ohnehin Null ergeben.

Der Anpassungstest von Kolmogorov ist im vorliegenden Fall wenig geeignet, es wird vielmehr ein quantitatives Vergleichsmaß benötigt.

Zur Bewertung der Übereinstimmung zwischen Verwendungspotential und zugehöriger Nutzung wird ein Überdeckungsgrad δ als derjenige prozentuale Anteil einer von einer Verteilung begrenzten Fläche definiert, der von der

anderen Verteilungsfläche überdeckt wird. Wegen der starken Streuung des Verwendungspotentials tritt beim Telefon mit $\sigma = 38,6$ % der relativ kleinste Wert auf. Mit Überdeckungsgraden zwischen 38 und 54 % wurde ein gutes Ergebnis erzielt, wie es gleichmäßiger nicht erwartet werden kann. Um den Einfluß der Verteilungsform zu eliminieren, könnte man die ermittelten Überdeckungsgrade noch mit den jeweils größten, bei Abweichung Null erreichbaren Werten σ_{\max} vergleichen, doch wird hierauf verzichtet.

8. Verwendungspotentiale zukünftiger Konfigurationen

Durch diesen Vergleich erfährt das entwickelte Bedarfsmodell eine Bestätigung; es haben sich zumindest keine Gesichtspunkte ergeben, die den in der Bedarfsanalyse getroffenen Annahmen widersprechen. Mit dieser Grundlage soll nun versucht werden, die Möglichkeiten und Grenzen technischer Verbesserungen der Kommunikationsmittel abzuschätzen. Dabei gehen wir von den Voraussetzungen aus, daß die Struktur des Bedarfsmodells und die Gewichtung der Bedarfsfaktoren im betrachteten zukünftigen Zeitraum von einigen Jahren weitgehend ihre Gültigkeit behalten.

Wenden wir uns zunächst dem Fernschreiber als dem Mittel mit dem bislang geringsten Nutzungsanteil zu, vgl. Bild 7. Die gegenwärtige Konfiguration der Kommunikationsmittel, für welche die Befragung durchgeführt wurde, wird als "Konfiguration I" bezeichnet. In der Konfiguration II sollen nun größtmögliche Verbesserungen für den Einsatz des Fernschreibers vorgenommen worden sein, vor allem eine dezentralisierte Aufstellung von Geräten in größerer Zahl. Man kann davon ausgehen, daß durch eine stärkere Benutzung dieses Kommunikationsmittels verschiedene psychologische Hürden in seiner Anwendung beseitigt werden. So wird im Laufe der Zeit sicherlich die Verständigung und die Förderung der Zusammenarbeit mit Hilfe dieses Mittels verbessert. Es wird angenommen, daß die Bedarfsfaktoren Nr. 1, 2,

4, 6, 7 und 8 z.T. beträchtlich vergrößert werden. Unvermeidliche Nachteile können sich allerdings in den Faktoren Nr. 3 und 9 auswirken:

Je häufiger Fernschreiben übermittelt werden, desto mehr werden sie zur Gewohnheit, d.h. desto stärker büßt die Kommunikation an Nachdruck zur Veranlassung einer Aktion ein. Da gleichzeitig möglicherweise mit einem Mangel an Bedienungspersonal gerechnet werden muß, wurde der Erfüllungsgrad für die Bequemlichkeit auf das bei Briefen festgestellte Niveau reduziert. Damit soll vor allem sichergestellt werden, daß der Brief als der vermutliche Hauptkonkurrent des Fernschreibens keine Benachteiligung erfährt.

Wie die Berechnung ergibt, kann unter diesen eher optimistischen Annahmen der Telex-Anteil von 6,1 % auf 14 % erhöht werden. Da im Bedarfsfaktor eine (realistische) Querbewertung zwischen den Erfüllungsgraden für die einzelnen Mittel stattfindet, s. Gl. (5) und (6), die einer Kopplung gleichkommt, wirkt sich die Anteils Kürzung nicht etwa gleichmäßig auf die drei Alternativmittel des Fernschreibers aus. Wie zu erwarten werden die Anteile von Telefon und Dienstreisen kaum, der des Briefes jedoch nennenswert reduziert.

In Bild 8 sollen Veränderungen studiert werden, die sich möglicherweise bei Einführung eines Bildtelefons ergeben. Ausgangszustand ist hier die Konfiguration II mit verbessertem Fernschreibereinsatz. Der Einfachheit halber wird das herkömmliche Telefon durch ein Bildtelefon ersetzt, also nicht der Wettbewerb zwischen beiden Varianten untersucht, Konfiguration III.

In Ermangelung genauer Kenntnisse über die Psychologie des Bildtelefonbetriebes wird angenommen, daß die Erfüllungsgrade der Bedarfsfaktoren 1, 2 und 3 erhebliche Verbesserungen erfahren und Größen erreichen, die nur noch von den einem persönlichen Gespräch (Dienstreise) zugeordneten Werten übertroffen werden. In den Bedarfsfaktoren 5, 6 und 9, nämlich bezüglich der Belästigung des Partners, der Effizienz und der Bequemlichkeit der

Durchführung, wird mit Verschlechterungen gerechnet. Der Anteil des Bildtelefons wächst gegenüber dem Telefon um 7,6 % auf 55,1 %, wodurch vor allem Brief und Dienstreise gewisse Abstriche erfahren.

Als Konfiguration IV sei noch die Variante betrachtet, daß der Einsatz des Bildtelefons mit nenneswerten Erschwernissen bei Dienstreisen zusammenfällt. Dabei wird angenommen, daß Dienstreisen zu diesem zukünftigen Zeitpunkt mit Verzögerungen verbunden sind, die dann als schwer zumutbar gelten, und daß die Effizienz solcher Reisen als relativ gering empfunden wird (Bedarfsfaktoren 6 und 7). Alle anderen Erfüllungsgrade für die Dienstreise werden nicht verändert, und die Verhältnisse bei allen anderen Kommunikationsmitteln bleiben gegenüber Konfiguration III konstant. Der Anteil der Dienstreisen am gesamten Kommunikationsaufkommen würde dann auf fast ein Drittel - nämlich auf 6,6 % - absinken, was angesichts der unverändert hohen Erfüllungsgrade Nr. 1, 2 und 3 auf den ersten Blick überrascht. Der Zuwachs käme relativ gleichmäßig dem Brief, dem Fernschreiben und dem Bildtelefon zugute. Man muß sich jedoch vor Augen halten, daß dieses Ergebnis nur unter der Annahme eines vollständigen Ersatzes des Telefons durch das Bildtelefon gilt. Bei Einführung des Bildtelefons wird dessen Anteil gegenüber dem Telefon jedoch zunächst sehr gering sein.

9. Abschließende Bemerkungen

Es kann wohlgemerkt nicht Zweck dieser Arbeit sein, eine Prognose über die Zukunft von Kommunikationsmitteln zu erstellen; hierzu wären detailliertere Analysen erforderlich. Die Beispiele in Abschnitt 8 sollen jedoch zeigen, wie man mit Hilfe eines überschaubaren Modells Ursachen und Wirkungen einzelner Veränderungen im Bedarfs- und Verwendungsgefüge eines "Marktes" untersuchen kann. Damit ist nicht zuletzt auch eine nützliche Grundlage

für weitere Diskussionen über dieses Thema geschaffen. Dieses Prinzip einer Bedarfsanalyse kann - wie eingangs erläutert - auch für Personenverkehrsstudien verschiedenster Art sowie für andere Untersuchungen angewandt werden, bei denen die Erfahrungen und Meinungen zahlreicher Personen zugrundegelegt werden sollen. Wünschenswert ist, daß man im Ablauf der Bedarfsanalyse eine Schleife zur Untersuchung auf Widerspruchsfreiheit, vgl. Bild 1, vorsieht. Der Wert einer solchen OR-Studie liegt - wie so häufig - weniger in ermittelten Prozentsätzen, als vielmehr in den gewonnenen Einsichten in die Struktur des Problems und in den Erkenntnissen, welche sich hieraus für den Planungs- und Entscheidungsprozeß ergeben können.

10. Literatur

- [1] H. M. Dathe: "Moderne Projektplanung in Technik und Wissenschaft. Modelle, Methoden und ihre Anwendungen." Carl Hanser Verlag, München, 1971

- [2] A. N. Sommers, F. F. Leimkuhler: " A Nondemographic Factor V/STOL Prediction Model." Paper presented to the ORSA National Meeting, Philadelphia, November 1968

- [3] A. N. Sommers: "Expanding Nondemographic Factors in Modal Split Models." Paper presented to the ORSA National Meeting, Miami, November 1969

- [4] P. R. Hofstätter: "Psychologie." Fischer-Lexikon Band 6, Frankfurt 1957

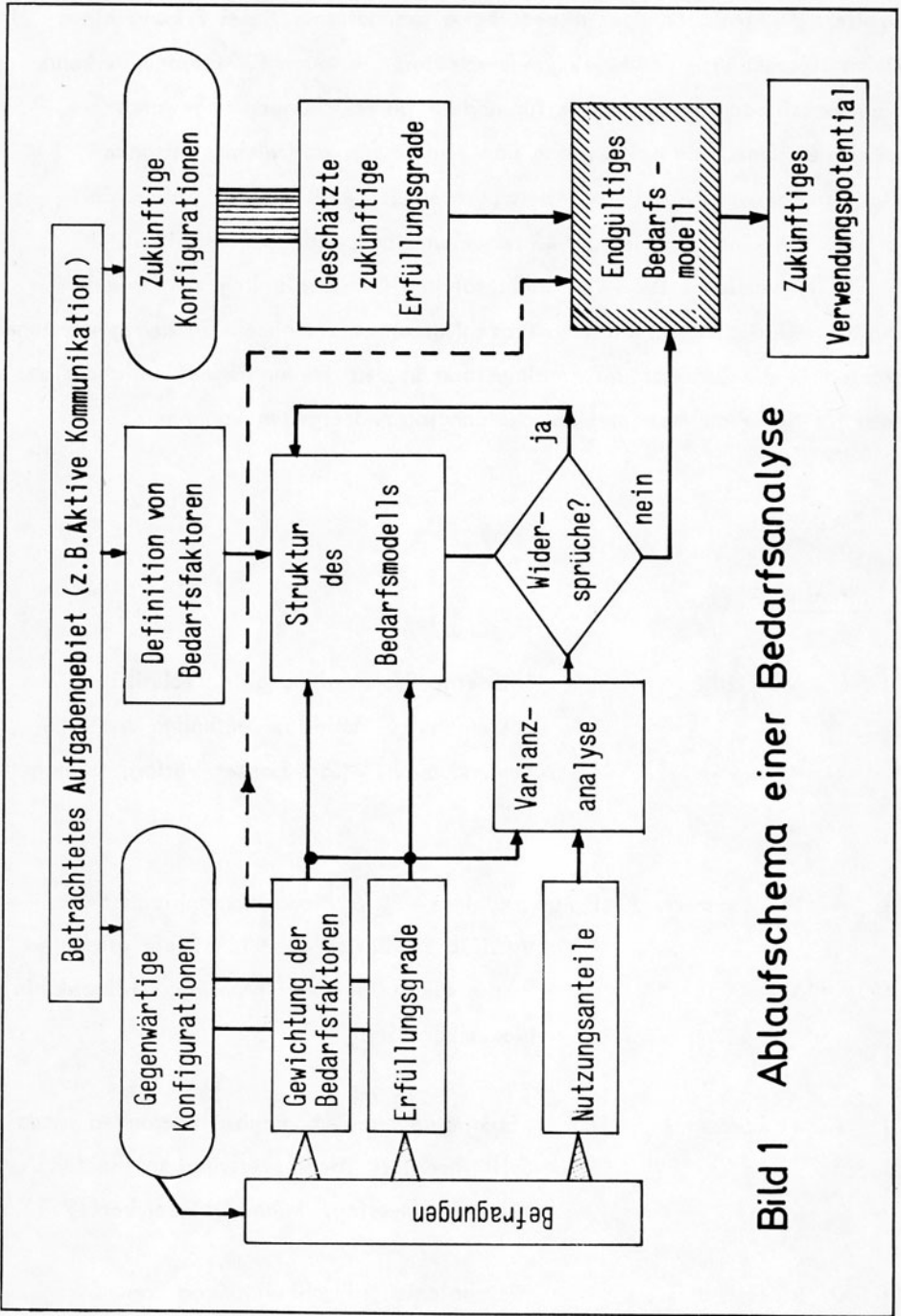


Bild 1 Ablaufschema einer Bedarfsanalyse

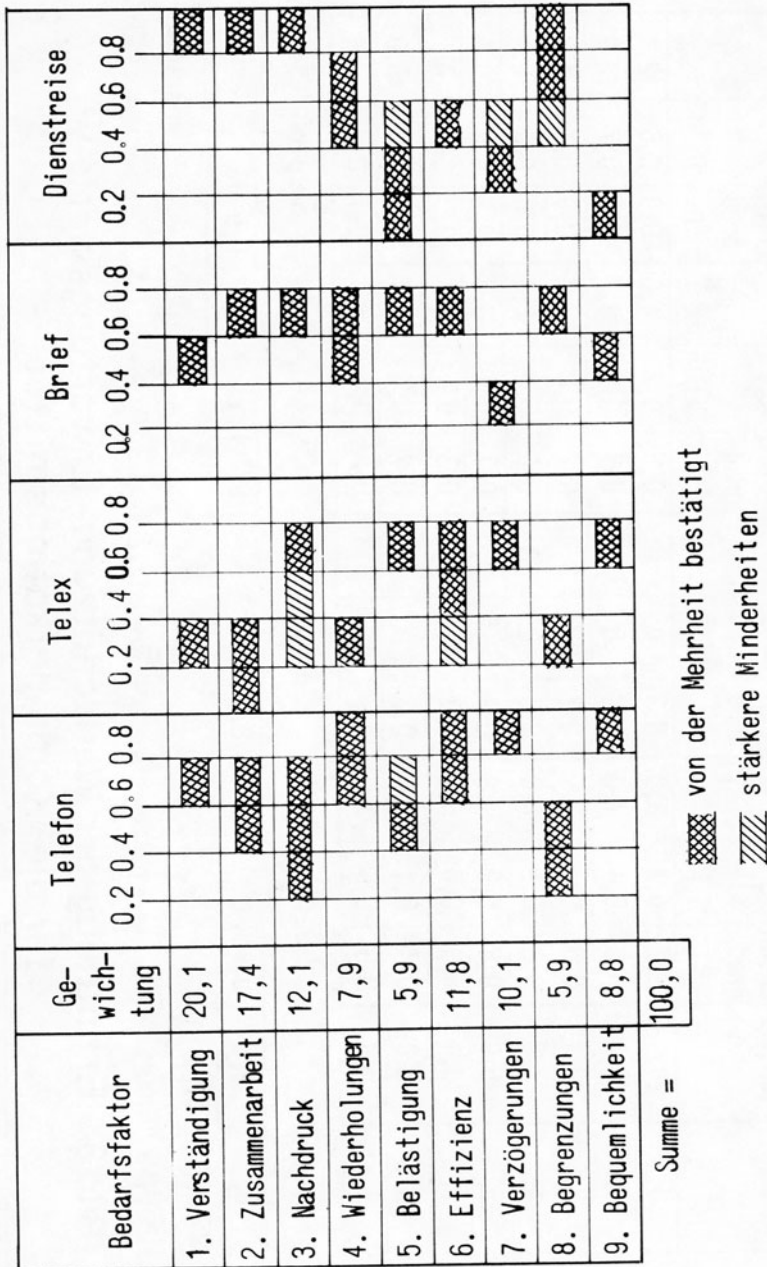


Bild 2 Bedarfsfaktoren für Kommunikationsmittel: Erfüllungsgrade und Gewichtungen (Befragungsergebnisse)

Bedarfs- faktor Nr.	Gewich- tung p	Telefon		Telex		Brief		Dienstreise	
		mittl. Erfüll- grad a_T	$\alpha_T =$ $a_T - a_0$	mittl. Erfüll- grad a_X	$\alpha_X =$ $a_X - a_0$	mittl. Erfüll- grad a_B	$\alpha_B =$ $a_B - a_0$	mittl. Erfüll- grad a_D	$\alpha_D =$ $a_D - a_0$
1.	20,1	0,7	0,4	0,3	0	0,5	0,2	0,9	0,6
2.	17,4	0,6	0,4	0,2	0	0,7	0,5	0,9	0,7
3.	12,1	0,5	0	0,7	0,2	0,7	0,2	0,9	0,4
4.	7,9	0,8	0,5	0,3	0	0,6	0,3	0,6	0,3
5.	5,9	0,5	0,3	0,7	0,5	0,7	0,5	0,2	0
7.	10,1	0,9	0,6	0,7	0,4	0,3	0	0,3	0
8.	5,9	0,4	0,1	0,3	0	0,7	0,4	0,8	0,5
9.	8,8	0,9	0,8	0,7	0,6	0,5	0,4	0,1	0

$$\sum p \cdot \alpha_T = 34,4$$
$$\sum p \cdot \alpha_X = 14,7$$
$$\sum p \cdot \alpha_B = 26,4$$
$$\sum p \cdot \alpha_D = 34,5$$

6.

~~11,8~~

0,8

$\times 0,8$

0,6

$\times 0,6$

0,7

$\times 0,7$

0,5

$\times 0,5$

$\sum W_i = 72,1$

$\sum \overline{W}_i = 100 \%$

$W_T = 27,5$

$\overline{W}_T = 38,1 \%$

$W_X = 8,8$

$\overline{W}_X = 12,2 \%$

$W_B = 18,5$

$\overline{W}_B = 25,7 \%$

$W_D = 17,3$

$\overline{W}_D = 24,0 \%$

Bild 3 Ermittlung der Wertschätzung - Erwartungswerte \overline{W}_i für die Kommunikationsmittel

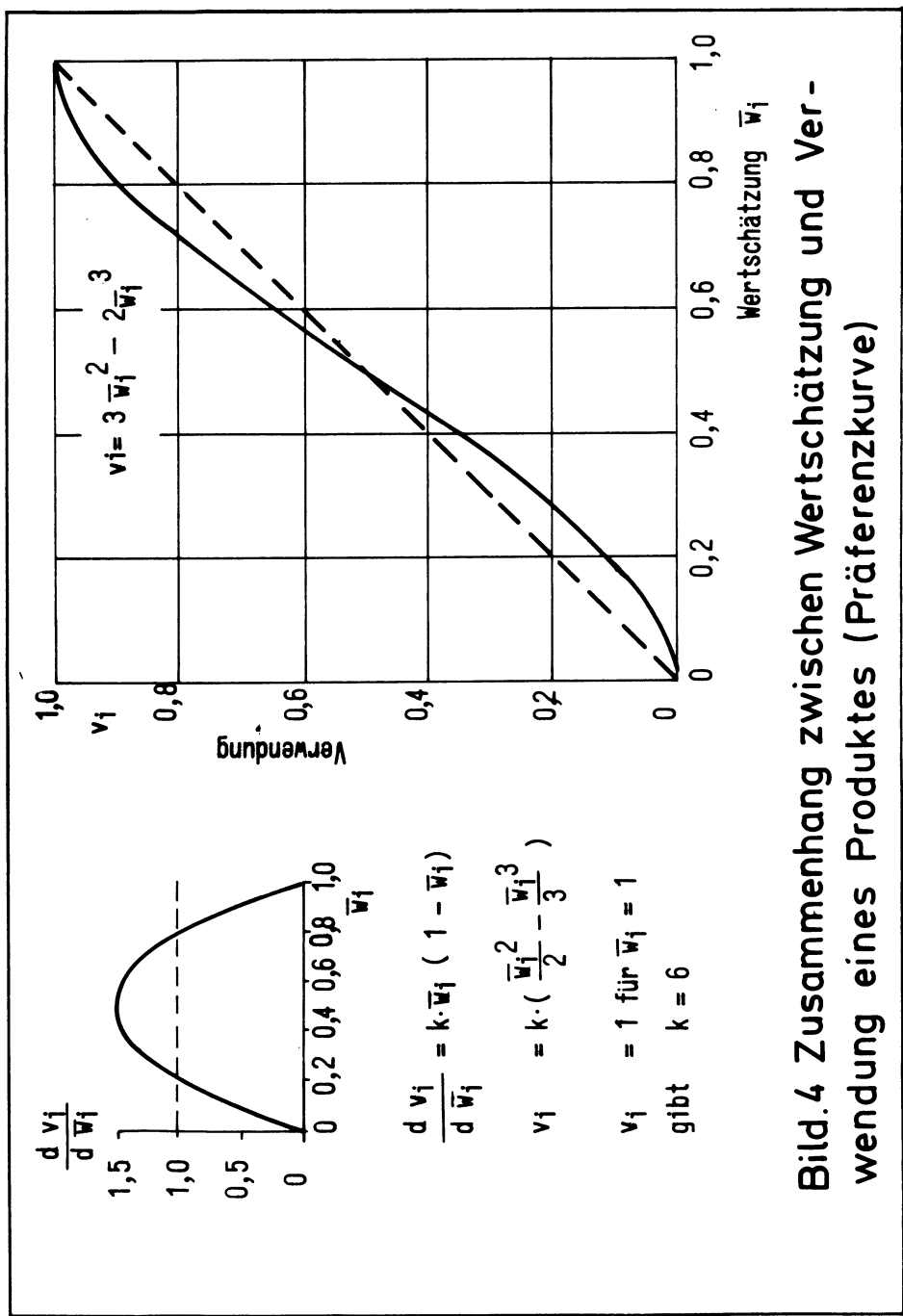


Bild.4 Zusammenhang zwischen Wertschätzung und Verwendung eines Produktes (Präferenzkurve)

Produkte $i = 1, 2, \dots, m$

Bedarfsfaktoren $j = 1, 2, \dots, n$

deren Gewichtungen p_j

und Erfüllungsgrade a_{ij}

$$\text{Bezugsgröße} \quad a_{0j} = \min_i [a_{ij}]$$

$$\text{Wertschätzung} \quad \bar{w}_i = \frac{a_{ie} \sum_j p_j (a_{ij} - a_{0j})}{\sum_i a_{ie} \sum_j p_j (a_{ij} - a_{0j})}$$

$$\text{Verwendungspotential} \quad \bar{v}_i = \frac{3 \bar{w}_i^2 - 2 \bar{w}_i^3}{\sum_i (3 \bar{w}_i^2 - 2 \bar{w}_i^3)}$$

Varianzanalyse :

$$\sigma^2(S_i) = \sum_j p_j \cdot \sigma^2(a_{ij}) \quad \text{wobei} \quad S_i = \sum_j p_j (a_{ij} - a_{0j})$$

$$\sigma^2(w_i) = \sigma^2(a_{ie}) \cdot \sigma^2(S_i) + a_{ie}^2 \cdot \sigma^2(S_i) + S_i^2 \cdot \sigma^2(a_{ie})$$

**Bild 5 Gleichungen des Bedarfsmodells
und der Varianzanalyse**

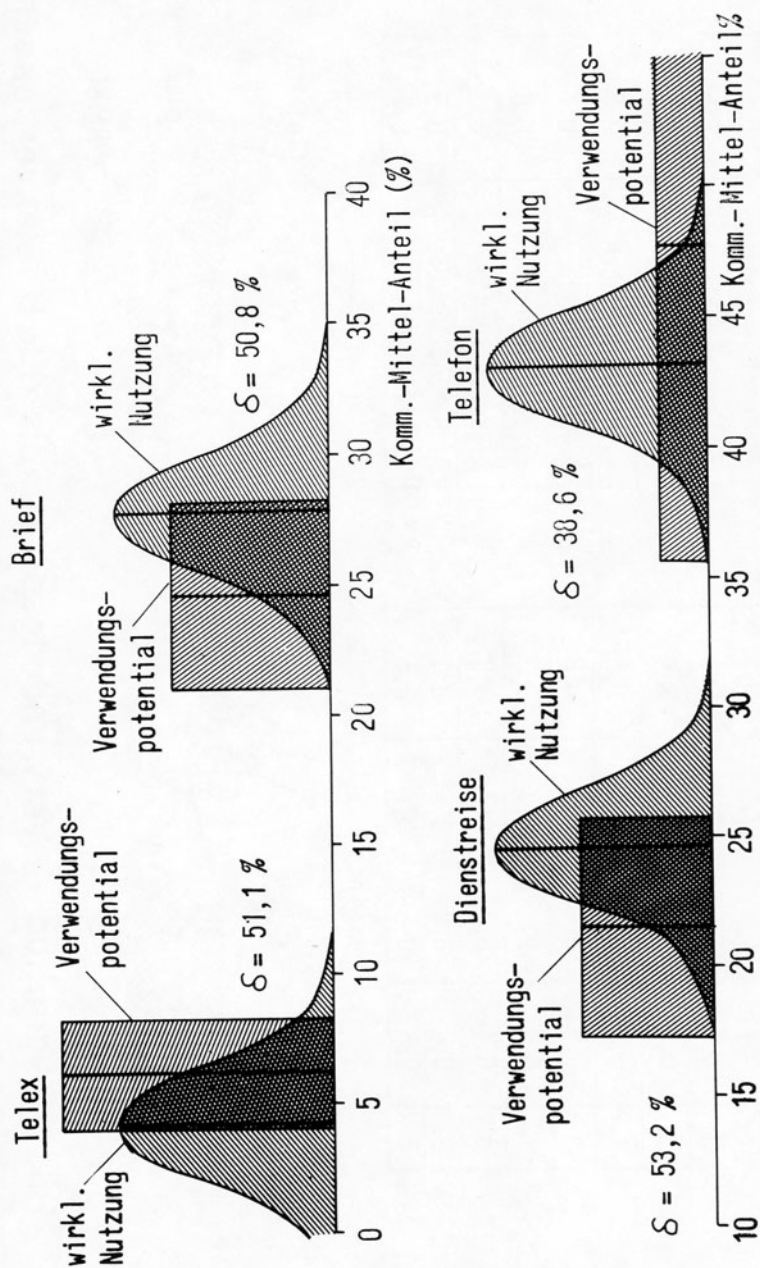


Bild6 Verwendungspotential und Nutzung von Kommunikationsmitteln (nach Befragung)

Gegenwärtige Konfiguration I

Nr.	Telefon a_T	Telex a_X	Brief a_B	D.-Reise a_D
1.	0,7	0,3	0,5	0,9
2.	0,6	0,2	0,7	0,9
3.	0,5	0,7	0,7	0,9
4.	0,8	0,3	0,6	0,6
5.	0,5	0,7	0,7	0,2
7.	0,9	0,7	0,3	0,3
8.	0,4	0,3	0,7	0,8
9.	0,9	0,7	0,5	0,1
6.	0,8	0,6	0,7	0,5

Zukünftige Konfiguration II

Bedarfsfaktor	Telex a_X
1. Verständigung	0,5
2. Zusammenarbeit	0,5
3. Nachdruck	0,6
4. Wiederholungen	0,5
5. Belästigung	0,7
7. Verzögerungen	0,8
8. Begrenzungen	0,5
9. Bequemlichkeit	0,5
6. Effizienz	0,7

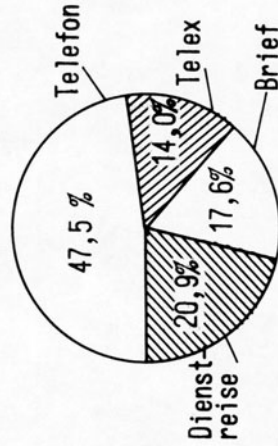
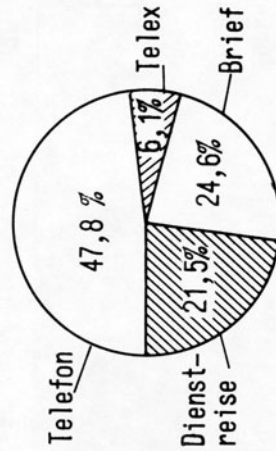
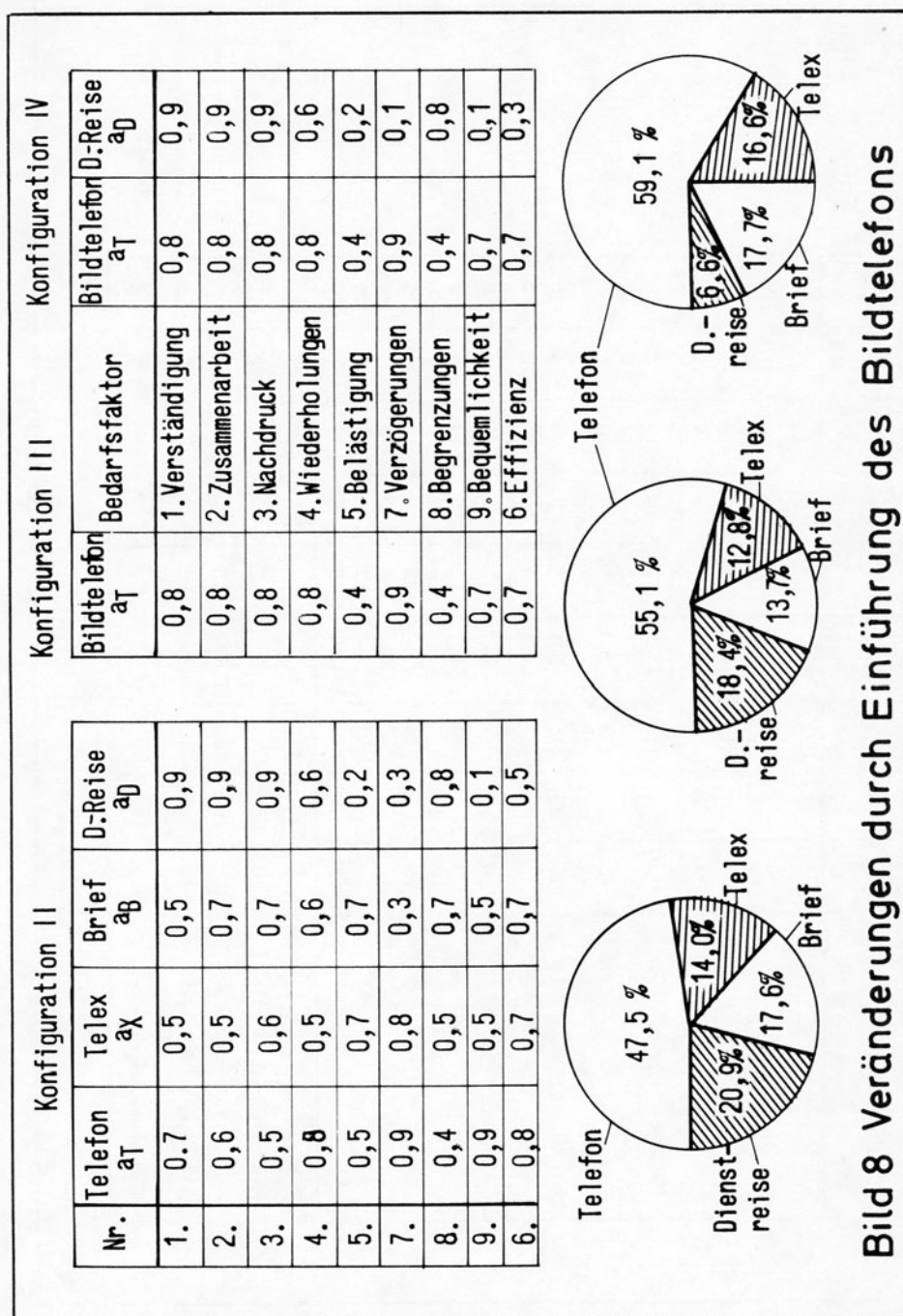


Bild 7 Variation d. Verwendungspotentials durch Telexverbesserg.



Bedarfsfaktor	j =	1	2	3	4	5	6	7	8	9
Vergabe von jeweils 90 Punkten durch 20 Befragte		15	10	5	10	5	15	10	5	15
		15	15	15	10	5	5	5	10	10
		18	18	18	5	4	9	5	4	9
		30	20	15	2	5	10	5	2	1
		18	16	8	10	4	6	14	12	2
		10	30	8	5	2	15	5	5	10
		15	5	15	15	5	10	10	5	10
		10	30	10	8	5	15	10	0	2
		40	20	5	2	1	2	10	4	6
		15	10	5	10	5	10	15	5	15
		20	15	15	5	5	10	5	5	10
		20	20	5	5	5	5	5	10	15
		16	18	14	4	12	8	10	2	6
		20	15	5	5	10	15	8	4	8
		20	15	10	10	5	10	10	5	5
		12	12	11	10	6	12	10	9	8
		15	15	15	5	10	10	10	5	5
		15	10	10	10	5	15	15	5	5
		20	10	10	10	10	2	15	10	8
		18	10	18	2	6	14	10	10	2
Punktsumme		362	314	217	143	107	211	182	107	157
kleinste vergeb. Punktzahl		10	5	5	2	1	2	5	0	1
mittlere Punktzahl		18,1	15,7	10,9	7,1	5,3	10,6	9,1	5,3	7,9
größte vergeb. Punktzahl		40	30	18	15	12	15	15	12	15
Gewichtung (Basis 100) P_j		20,1	17,4	12,1	7,9	5,9	11,8	10,1	5,9	8,8

Tabelle 1
Befragungsergebnis
für die Gewichtung
der Bedarfs-
faktoren

Befragung Nr.	Telefon	Telex	Brief	Dienst- reise	(c-b)	$\sigma^2 = \frac{(\sum c - b)^2}{12}$
1	70	0	10	20	10	8,3
2	48,4	3,3	16,1	32,2	0,1	0
3	30	0	20	50	10	8,3
4	50	0	30	20	10	8,3
5	50	5	30	15	5	2,1
6	62,5	0	12,5	25	2,5	0,5
7	40	10	30	20	10	8,3
8	30	0	20	50	10	8,3
9	20	0	70	10	10	8,3
10	60	5	15	20	5	2,1
11	35	20	25	20	5	2,1
12	45	10	40	5	5	2,1
13	25	5	50	20	5	2,1
14	60	5	30	5	5	2,1
15	36	4	32	28	2	0,3
16	10	0	20	70	10	8,3
17	60	5	25	10	5	2,1
18	50	0	25	25	5	2,1
19	40	10	30	20	10	8,3
Summe	821,9	82,3	530,6	465,2	$\sigma_u^2 =$	84,0
$u_i =$	43,2 %	4,3 %	27,9 %	24,6 %		

Tabelle 3 Befragungsergebnis für die Nutzungsanteile und zugehörige Streuung

Tabelle 4 Varianzanalyse für die ermittelten Wertschätzungsgrößen

Nr.	Gewichtung p_j	Telefon		Telex		Brief		Dienstreise	
		$(c-b)_{\alpha_T}$	$p_j \cdot \sigma^2(\alpha_T)$	$(c-b)_{\alpha_x}$	$p_j \cdot \sigma^2(\alpha_x)$	$(c-b)_{\alpha_B}$	$p_j \cdot \sigma^2(\alpha_B)$	$(c-b)_{\alpha_D}$	$p_j \cdot \sigma^2(\alpha_D)$
1.	20,1	0	0	0	0	0	0	0	0
2.	17,4	0,4	0,231	0	0	0,2	0,058	0,2	0,058
3.	12,1	0	0	0,4	0,161	0,4	0,161	0,4	0,161
4.	7,9	0,2	0,026	0	0	0,2	0,026	0,2	0,026
5.	5,9	0,2	0,019	0,2	0,019	0,2	0,019	0	0
7.	10,1	0	0	0	0	0	0	0	0
8.	5,9	0,2	0,019	0	0	0	0	0,2	0,019
9.	8,8	0	0	0	0	0	0	0	0
$\sum_j p_j \cdot \sigma^2(\alpha_{ij}) =$		$\sigma^2(S'_T) = 0,295$		$\sigma^2(S'_x) = 0,180$		$\sigma^2(S'_B) = 0,264$		$\sigma^2(S'_D) = 0,26$	
6.	1	0,4	0,0133	0,4	0,0133	0,2	0,0033	0,2	0,0033

$$\sigma^2(w_T) = 15,99 \quad \sigma^2(w_x) = 2,92 \quad \sigma^2(w_B) = 2,43 \quad \sigma^2(w_D) = 3,99$$

Marginal Utility in the Economization of Power Series

von K. Seiler, McLean

ABSTRACT

This paper deals with a comparison of an ordinary power series in which the error tends to be large at the ends of the interval and small in the middle versus lower ordered Tschebyshev polynomials which result in minimum error at the ends of the interval. It is suggested that extrapolation error would be minimized through the use of Tschebyshev polynomials which converge more rapidly (have higher marginal utility per term) than any other set of orthogonal polynomials. It is further suggested that where possible a tradeoff be made of the marginal utility of adding another datum point to the sample versus adding another term to the power series, in order to achieve an "optimal" balance between these two contributors to error reduction.

I. INTRODUCTION

A. Purpose

The first purpose of this paper is to show that in the development of a CER (cost estimating relation) which is non-linear and in the form of an ordinary power series, such as a Taylor series, that for purposes of extrapolation it may be more accurate to use a lower ordered Tschebyshev polynomial. The second purpose is to call attention to the possible trade-off of datum points in the sample for terms in the power series in order to achieve some "optimal" equilibrium between these two contributors to error reduction.

B. Background

It frequently occurs in the development of CER's that the cost function increases at an increasing rate resulting in a non-linear expression in the form of an ordinary power series. It is a characteristic of ordinary power series that the error tends to be large at the extremes of the interval under investigation and a minimum in the middle of the interval. Since extrapolation involves the extension of the extremes of an interval it follows that a reduction in the error at the interval extremes should result in less error in extrapolation. By employing a Tschebyshev polynomial approximation of the ordinary power series, the extremes error will be reduced, but the average square error will be larger.

II. ANALYSIS

A. Identification of Variables

In the study of Dimensional Analysis it is stressed that the identification of the independent variables in any phenomenon under investigation requires considerable philosophical insight, experimentation, and testing. The cost analyst should consult with design engineering for assistance in this process if necessary. Obviously, the rule of parsimony should prevail--- that the number of independent variables be no more than required to achieve a given level of accuracy.

B. Power Series Form of CER

There are numerous forms of ordinary power series expressions which could be used to describe various non-linear cost estimating relations. Regardless of the particular form of a specific power series expression, these expressions are all expansions of a basic formula. Consider a truncated power series of a function such as:

$$(1) \quad f(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$$

in an interval $(-1 \leq x \leq 1)$ where x is the dimensionless product of a number of independent variables.

As stated in the introduction, the error is a maximum at the ends of the interval and a minimum in the middle.

C. Economization Process

The first step in the process of economization uses the following table of the trigonometric equivalents of the powers of x to convert the power series to an expansion in Tschchebyshev orthogonal polynomials:

$$\begin{aligned}
 1 &= T_0 \\
 x &= T_1 \\
 x^2 &= 1/2 (T_0 + T_2) \\
 (2) \quad x^3 &= 1/4 (3T_1 + T_3) \\
 x^4 &= 1/8 (3T_0 + 4T_2 + T_4) \\
 x^5 &= 1/16 (10T_1 + 5T_3 + T_5)
 \end{aligned}$$

resulting in:

$$(3) \quad f(x) = b_0 + b_1 T_1(x) + b_2 T_2(x) + \dots + b_n T_n(x)$$

For a large group of functions the expansion in Tschebyshev polynomials converges more rapidly than the expansion in any other set of orthogonal polynomials. (See Reference 2). Stated more explicitly, the b_k in equation (3) should become small more rapidly than the a_k of equation (1).

D. Example

To illustrate the economization process numerically let's consider the following simple power series:

$$(4) \quad y = \ln(1+x) \approx x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5}$$

Using the table of equations (2), equation (x) becomes

$$\begin{aligned}
 (5) \quad y &= T_1 - 1/4(T_0 + T_2) + 1/12(3T_1 + T_3) - 1/32(3T_0 + 4T_2 + T_4) \\
 &\quad + 1/80(10T_1 + 5T_3 + T_5) \\
 y &= 11/32T_0 + 11/8T_1(x) - 3/8T_2(x) + 7/48T_3(x) - 1/32T_4(x) + 1/80T_5(x)
 \end{aligned}$$

In the interval $(0 \leq x \leq 1)$, deleting the last term of the power series in equation (4) produces a change of $1/5$ (if $x=1$), whereas deleting the last three terms in equation (5) produces a change of approximately the same magnitude:

$$7/48 + 1/32 + 1/80 = 91/480 < .19$$

Clearly, the Tschebyshev polynomial converges more rapidly (has higher marginal utility per term) than the ordinary power series.

Knowing that $T_k(x) \leq 1$ for $(-1 \leq x \leq 1)$ equation (5) can be economized to:

$$(6) \quad y = \ln(1+x) \approx -11/32T_0 + 11/8T_1(x) = 3/8T_2(x)$$

The economized Tschebyshev expansion represented by equation (6) may now be converted back to an economized polynomial using the converse of the table of equations in (2):

$$T_0 = 1$$

$$T_1 = x$$

$$T_2 = 2x^2 - 1$$

$$T_3 = 4x^3 - 3x$$

$$T_4 = 8x^4 - 8x^2 + 1$$

$$T_5 = 16x^5 - 20x^3 + 5x$$

Therefore, equation (5) in economized form is reduced to:

$$y = \ln(1+x) \approx -11/32 + 11/8(x) - 3/8(2x^2 - 1) \approx 1/32 + 11x/8 - 3x^2/4$$

E. Tradeoff of Datum Points in Sample versus Terms in Power Series

The rationale underlying this tradeoff procedure rests on utility theory. Graphically, the process may be represented by a convex (decreasing marginal rate of substitution) curve on a utility map as follows:

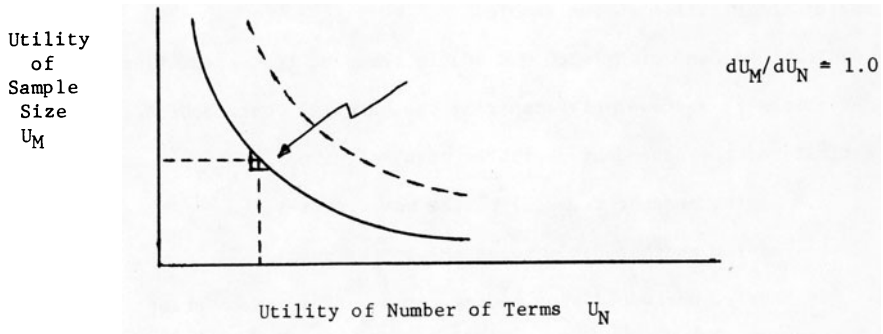


Figure 1

By specifying (ideally) that the marginal contributors to the estimating process be of equal magnitude, there is established thereby a unique solution which defines the "optimum" mix of the two contributors. This optimum occurs if and only if $\Delta U_M = \Delta U_N$ as shown by the dotted coordinate intercepts in Figure 1.

There are three fundamental constraints assumed in this optimization technique:

(1) The sample size is small (≤ 30). Naturally, for large samples, this technique will cause the number of terms in the series to be impractically large; however, for small samples the marginal contribution to the estimate of the Mth sample becomes significant and at some point equals the marginal contribution to the estimate of an Nth term in the power series.

(2) Resources are fixed re-sampling.

Removal of this constrain merely displaces the utility function; for example, in Figure 1, the dotted isoquant would be at a higher level of resources.

(3) The cost of adding the Nth term to the power series = the cost of adding the Mth item to the sample.

The "optimum" of N (when $M \leq 30$) in terms of best predictions may be determined by successively comparing the marginal contribution to the estimate of the dependent variable (cost) of:

a - adding another term (N) to the power series.

b.- adding another datum point (M) to the sample.

In theory, one would stop adding terms to the power series when the marginal return from the Nth term in the series = the marginal return from the Mth item in the sample. In practice, identify would be rarely achieved; therefore, it is suggested for practical purposes that the iteration cease whenever the marginal contribution of the Nth term becomes \leq the marginal contribution of the Mth item. The method then is one of a converging series of successive tradeoffs of estimating power gained by adding terms to the power series versus estimating power gained by adding datum points to the sample.

REFERENCES

1. Conte, S. D., "Elementary Numerical Analysis," McGraw-Hill, 1965, Sec. 3.8.
2. Hamming, R. W., "Numerical Methods for Scientists and Engineers," McGraw-Hill, 1962, Chap. 19.
3. Householder, Alston S., "Principles of Numerical Analysis," McGraw-Hill, 1953, Sec. 6.2.

Spektralanalytische Untersuchungen von Aktienkursentwicklungen
von W. Stier, Bochum

Die zeitliche Entwicklung von Aktienkursen war in den letzten Jahren häufig Gegenstand statistischer Analysen. ¹⁾ Die bekanntesten von ihnen sind wohl die Untersuchungen von C.W.J. Granger und O. Morgenstern, Spektralanalysen von Kursnotierungen der New Yorker Börse. ²⁾ Granger/Morgenstern diskutierten die Auffassung früherer Autoren, daß sich die Entwicklung von Aktienkursen durch ein einfaches Modell beschreiben läßt, das in der Theorie der stochastischen Prozesse als "random-walk"-oder "Irrfahrt-Modell" bekannt ist. ³⁾ Wenn X_t der Kurs eines Papiers zum Zeitpunkt t ist, dann gilt bei diesem Modell

$$X_t - X_{t-1} = \epsilon_t$$

mit $E(\epsilon_t) = 0$ und $E(\epsilon_t \epsilon_{t-s}) = 0$ für $s \neq 0$. Der Prozeß $\{\epsilon_t\}$, $t = 1, 2, \dots$, wird als weißes Rauschen bezeichnet.

Neben der Frage, ob dieses einfache Modell hinreichend für die Erklärung der Kursentwicklung an der New Yorker Börse ist, interessieren sich Granger/Morgenstern u.a. für folgende Probleme: ⁴⁾

-
- 1) Es sei hier stellvertretend auf folgende Arbeiten hingewiesen:
M.G. Kendall, The Analysis of Economic Time Series - Part I: "Prices", I. Royal Stat. Soc. (Series A), Vol. 96, 1953 S. 11 - 25
S. Alexander, Price Movements in Speculative Markets: Trends or Random Walks. Industrial Management Review. Vo. 2 (1961) S. 7 - 26
P.H. Cootner, Stock Prices, Random vs. Systematic Changes. Industrial Management Review, Vol. 3 (1962) S. 24 - 45
 - 2) C.W.J. Granger und O. Morgenstern, Spectral Analysis of New York Stock Market Prices, Kyklos No. 16 (1963), S. 1-27. Derselbe Artikel findet sich auch in: The Random Character of Stock Market Prices, herausgegeben von P.H. Cootner, S. 162 ff.
 - 3) Nach Auffassung von Granger/Morgenstern soll diese Bezeichnung in diesem Zusammenhang nicht ganz zutreffend sein, da hier ungleiche Schrittweiten zugelassen seien. Darauf kommt es jedoch nicht an, da es "Random-Walks" gibt, bei denen die Schrittweiten ungleich sind. Siehe W.Feller: An Introduction into Probability Theory and its Application, New York 1950, S. 330

- 1) Läßt sich der US-Konjunkturzyklus von 40 Monaten in den Kursen nachweisen?
- 2) Unterliegen die Kurse saisonalen Schwankungen?
- 3) Sind lead-lag-Beziehungen zwischen den Kursen verschiedener Papiere festzustellen?

Um die random-walk Hypothese zu testen, wurden die Spektren der ersten Differenzen verschiedener Aktienindizes geschätzt.¹⁾ Sie verliefen alle im gesamten Frequenzbereich sehr flach. Nur wenige Schätzwerte lagen außerhalb eines 95 %-Konfidenzintervalls. Damit schien die random-walk Hypothese die Kursentwicklung an der New Yorker Börse befriedigend zu erklären. Nach Granger/Morgenstern ist dabei aber folgendes zu bedenken: Betrachtet man die beiden Prozesse

$$X_t = X_{t-1} + \varepsilon_t$$

und

$$Y_t = X_t - a \cos \omega t$$

und bildet ihre ersten Differenzen, so sind diese kaum voneinander zu unterscheiden, wenn ω klein ist, d.h., wenn Y_t eine periodische Komponente mit langer Periode enthält. Das einfache random-walk Modell sagt somit nichts über möglicherweise vorhandene langfristige Komponenten eines zufälligen Prozesses aus. Deswegen schätzen Granger/Morgenstern auch die Spektren der Originaldaten und kommen zu folgendem Resultat: im Mittel- und Hochfrequenzbereich wird das random-walk Modell bestätigt, nicht dagegen im Niederfrequenzbereich. Die geschätzten Spektren weisen auf eine vorhandene Trend/Zykluskomponente hin. Interessant ist jedoch, daß sich der US-Konjunkturzyklus, wenn überhaupt, nur

1) In einer späteren Arbeit haben Granger/Morgenstern zusammen mit Godfrey auch Tageskurse untersucht und dabei ihre vorherigen Ergebnisse bestätigt gefunden. Jedoch wurden die möglicherweise vorhandenen Beziehungen zwischen den Tageskursen einzelner Papiere nicht untersucht. Vgl. M.D. Godfrey, C.W.J. Granger und O. Morgenstern, The Random-Walk Hypothesis of Stock Market Behavior, Kyklos No. 17 (1964), S. 1 - 30

sehr schwach nachweisen läßt. Auch läßt sich keine signifikante Saisonkomponente feststellen, obwohl gerade diese nach landläufiger Meinung vorhanden sein müßte. Die geschätzten Kohärenzen und Phasen lassen keine lead-lag-Beziehungen zwischen den verschiedenen Indizes erkennen. Granger/Morgenstern sind der Auffassung, daß die Verhältnisse an anderen Börsen ähnlich sind, halten jedoch entsprechende Untersuchungen für erforderlich. Es soll nun über einige Ergebnisse einer entsprechenden Untersuchung für die BRD berichtet werden. Methodisch basiert die vorliegende Untersuchung auf der Theorie der Spektralanalyse stationärer stochastischer Prozesse.

Unter einem stochastischen Prozess versteht man folgendes: Gegeben ist ein Wahrscheinlichkeitsfeld (Ω, σ, P) , wobei Ω die Menge der Elementarereignisse, σ ein Ereignisring von Ω und P ein darauf definiertes Wahrscheinlichkeitsfunktional ist. Ordnet man jedem $\omega \in \Omega$ eine zeitabhängige Funktion

$$X(t, \omega), \quad \omega \in \Omega, \quad t \in T$$

zu, dann erhält man eine Familie von Funktionen oder einen stochastischen Prozeß. Ein stochastischer Prozeß ist somit eine Funktion der beiden Variablen ω und t . Für ein bestimmtes $t_1 \in T$ ist $X(t_1, \omega)$ eine auf σ meßbare Funktion, d.h. eine zufällige Variable und für ein bestimmtes $\omega_1 \in \Omega$ ist $X(t, \omega_1)$ eine Funktion der Zeit, d.h. eine sog. Realisierung des betrachteten stochastischen Prozesses. Sind z.B. die Elemente ω der Menge Ω die Wertpapiere gewisser Firmen, dann bedeutet $X(t, \omega_1)$ die zeitliche Kursentwicklung eines bestimmten Papiers. Die Spektralanalyse deutet somit Zeitreihen als Realisationen von stochastischen Prozessen, genauer: als Realisationen von diskreten, kovarianzstationären stochastischen Prozessen. Bei diesen Prozessen ist T eine abzählbare Menge und es gilt

$$R(t, t + \tau) = R(\tau) = E[X(t)X(t + \tau)] - \mu^2$$

mit $E X(t) = \mu$ ^{1), 2)}

1) Die Abhängigkeit der Funktion $X(t)$ von ω wird im folgenden nicht explizit zum Ausdruck gebracht.

2) Der Einfachheit halber sei im folgenden $\mu = 0$ gesetzt.

d.h. ihre Autokovarianzfunktion hängt nicht von t ab. Nach dem Satz von Chintchin¹⁾ läßt sich die Autokovarianzfunktion eines diskreten, autokovarianzstationären Prozesses in der Form

$$R(\tau) = \int_{-\pi}^{\pi} e^{i\tau\omega} dF(\omega) \quad (1)$$

darstellen. Dabei ist $F(\omega)$ das sogenannte kumulierte Spektrum des Prozesses und ω seine Frequenzen. Enthält der Prozeß keine streng periodischen Komponenten, dann ist $F(\omega)$ stetig differenzierbar und das sogenannte Spektrum

$$f(\omega) = \frac{dF(\omega)}{d\omega}$$

existiert. Durch Inversion von (1) ergibt sich für das Spektrum

$$f(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} R(\tau) e^{-i\tau\omega}$$

Für reelle Prozesse erhält man dafür

$$f(\omega) = \frac{1}{2\pi} \left\{ R(0) + 2 \sum_{\tau=1}^{\infty} R(\tau) \cos \tau\omega \right\} \quad (2)$$

da

$$R(\tau) = R(-\tau)$$

Autokovarianzfunktion und Spektrum sind fouriertransformierte Paare und enthalten dieselbe Infor-

1) A.I. Chintchin: Korrelationstheorie stationärer stochastischer Prozesse, Math. Annalen 109, 1934

mation über den zugrunde liegenden Prozeß. Während die Autokovarianzfunktion den Prozeß im Zeitbereich beschreibt, liefert das Spektrum eine Beschreibung im Frequenzbereich. Es ist jedoch schwierig, aus einer geschätzten Autokovarianzfunktion herauszufinden, welche Komponenten (z.B. saisonaler oder zyklischer Art) in welchem Umfang zur Erklärung der gesamten Varianz eines Prozesses beitragen. Hinzu kommt noch, daß die Schätzwerte in der Regel selbst stark korreliert sind. Diese Art der Information bietet in einfacher Weise das Spektrum, denn aus (1) folgt

$$R(0) = \text{Var } X(t) = \int_{-\pi}^{\pi} f(\omega) d\omega$$

d.h. die gesamte Varianz des Prozesses wird auf einzelne Frequenzen verteilt und es läßt sich an Hand des Spektrums feststellen, welche Komponenten die gesamte Variation des Prozesses bestimmen.

Eine naheliegende Schätzfunktion für das Spektrum bei Vorliegen von N Stichprobenwerten ist das sog. Periodogramm

$$P(\omega) = \frac{1}{2\pi N} \left| \sum_{t=1}^N X(t) e^{i\omega t} \right|^2 \quad (3)$$

Es läßt sich zeigen, daß gilt

$$P(\omega) = \frac{1}{2\pi} \left\{ \hat{R}_0 + \frac{1}{2} \sum_{\tau=1}^{N-1} \hat{R}(\tau) \cos \tau \omega \right\} \quad (4)$$

wobei

$$\hat{R}(\tau) = \frac{1}{N} \sum_{t=1}^{N-\tau} X(t) X(t + \tau)$$

eine konsistente Schätzfunktion für $R(\tau)$ ist. Bei Schätzfunktion (3) werden die Werte der vorliegenden Zeitreihe direkt verwendet (direkte Schätzmethode), während bei der Schätzfunktion (4) erst ihre Autokovarianzen berechnet wer-

den müssen (indirekte Schätzmethode). Beide Schätzfunktionen sind asymptotisch erwartungstreu, jedoch nicht konsistent. Konsistente Schätzfunktionen erhält man aus (3) und (4) durch Einführung von sog. Gewichtsfunktionen. Benützt man Funktion (3), so kann man vom Algorithmus der schnellen Fourier-Transformation Gebrauch machen. Die folgenden Spektren wurden mit der indirekten Methode geschätzt, wobei die Gewichtsfunktion von Tukey-Hanning verwendet wurde. Die Schätzfunktion lautet in diesem Fall

$$\hat{f}(\omega) = \frac{1}{2\pi} \left\{ \hat{R}(0) + 2 \sum_{\tau=1}^m \hat{R}(\tau) K\left(\frac{\tau}{m}\right) \cos \tau \omega \right\}.$$

m ist die Anzahl der verwendeten lags und die Gewichtsfunktion lautet

$$K(x) = \begin{cases} \frac{1}{2} (1 + \cos \pi x) & \text{für } |x| < 1 \\ 0 & \text{für } |x| \geq 1 \end{cases}$$

Will man nun mit Hilfe der Spektralanalyse die Frage klären, ob und gegebenenfalls welche Beziehungen zwischen den Kursentwicklungen verschiedener Papiere bestehen, so unterstellt man, die Kurse seien Realisationen eines reellen, bivariaten, stationären, stochastischen Prozesses. Ein bivariater stochastischer Prozeß $\{X(t), Y(t)\}$ ist kovarianzstationär, wenn gilt

$$E[X(t)] = \mu_x, \quad E[X(t) - \mu_x]^2 = \sigma_x^2$$

$$R_{XX}(t, t+\tau) = R_{XX}(\tau) = E[(X(t) - \mu_x)(X(t+\tau) - \mu_x)]$$

$$E[Y(t)] = \mu_y, \quad E[Y(t) - \mu_y]^2 = \sigma_y^2$$

$$R_{YY}(t, t+\tau) = R_{YY}(\tau) = E[(Y(t) - \mu_y)(Y(t+\tau) - \mu_y)]$$

$$R_{XY}(t, t+\tau) = R_{XY}(\tau) = E[(X(t) - \mu_x)(Y(t+\tau) - \mu_y)]$$

$R_{XY}(\tau)$ ist dabei die Kreuz-Kovarianzfunktion. Die Kovarianzfunktionen eines reellen bivariaten Prozesses lassen sich wie folgt darstellen

$$R_{XX}(\tau) = 2 \int_0^{\pi} \cos \tau \omega \, dF_X(\omega)$$

$$R_{YY}(\tau) = 2 \int_0^{\pi} \cos \tau \omega \, dF_Y(\omega)$$

$$R_{XY}(\tau) = 2 \int_0^{\pi} \cos \tau \omega \, dC(\omega) - 2 \int_0^{\pi} \sin \tau \omega \, dQ(\omega).$$

Enthält der Prozeß keine streng periodischen Komponenten, dann sind die Funktionen $F_X(\omega)$, $F_Y(\omega)$, $C(\omega)$, $Q(\omega)$ stetig differenzierbar und es gilt

$$R_{XX}(\tau) = 2 \int_0^{\pi} \cos \tau \omega \, f_X(\omega) \, d\omega$$

$$R_{YY}(\tau) = 2 \int_0^{\pi} \cos \tau \omega \, f_Y(\omega) \, d\omega$$

$$R_{XY}(\tau) = 2 \int_0^{\pi} \cos \tau \omega \, c(\omega) \, d\omega - 2 \int_0^{\pi} \sin \tau \omega \, q(\omega) \, d\omega$$

$$= \int_{-\pi}^{\pi} e^{i\tau\omega} C_r(\omega) \, d\omega.$$

Dabei sind $f_X(\omega)$ und $f_Y(\omega)$ die Spektren der Prozesse $\{X(t)\}$ und $\{Y(t)\}$ und

$$C_r(\omega) = c(\omega) + iq(\omega)$$

ist das Kreuzspektrum zwischen $\{X(t)\}$ und $\{Y(t)\}$. $c(\omega)$ heißt Kospektrum und $q(\omega)$ Quadratspektrum. Diese Funktionen genügen der Kohärenz-Ungleichung

$$c^2(\omega) + q^2(\omega) \leq f_X(\omega) f_Y(\omega) .$$

Durch Inversion obiger Beziehungen erhält man für die Spektren sowie für Ko- und Quadratspektrum

$$f_X(\omega) = \frac{1}{2\pi} \{R_{XX}(0) + 2 \sum_{\tau=1}^{\infty} R_{XX}(\tau) \cos \tau\omega\}$$

$$f_Y(\omega) = \frac{1}{2\pi} \{R_{YY}(0) + 2 \sum_{\tau=1}^{\infty} R_{YY}(\tau) \cos \tau\omega\}$$

$$c(\omega) = \frac{1}{2\pi} R_{XY}(0) + \frac{1}{\pi} \sum_{\tau=1}^{\infty} (R_{XY}(\tau) + R_{YX}(\tau)) \cos \tau\omega$$

$$q(\omega) = \frac{1}{\pi} \sum_{\tau=1}^{\infty} (R_{XY}(\tau) - R_{YX}(\tau)) \sin \tau\omega .$$

Ein Maß für den linearen Zusammenhang zwischen den Frequenzkomponenten der beiden Prozesse ist die Kohärenzfunktion

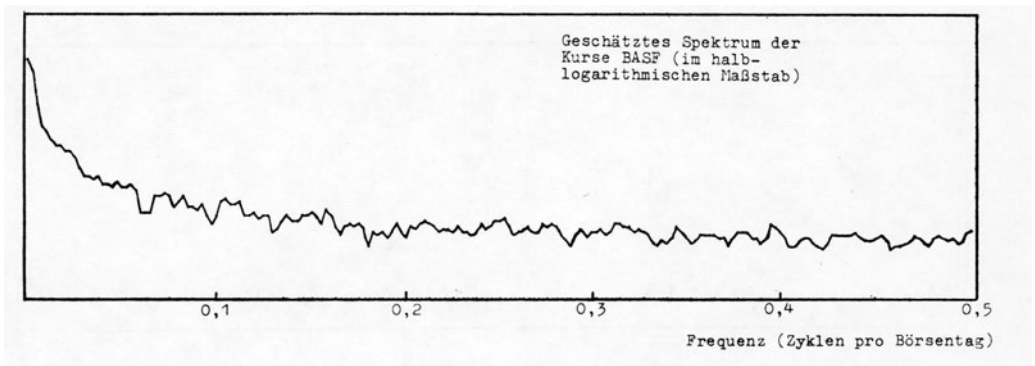
$$C(\omega) = \frac{c^2(\omega) + q^2(\omega)}{f_X(\omega) f_Y(\omega)}$$

mit $0 \leq C(\omega) \leq 1$. $C(\omega)$ ist analog zum Determinationskoeffizienten zu interpretieren. Ein Maß für die Phasendifferenz entsprechender Frequenzkomponenten liefert die Phasenfunktion

$$\psi(\omega) = \arctan \left(\frac{q(\omega)}{c(\omega)} \right)$$

Kohärenz und Phasenfunktion charakterisieren vollständig die Beziehungen zwischen zwei stationären Prozessen. Konsistente Schätzfunktionen für Ko- und Quadratspektrum erhält man analog zum unvariablen Fall durch Einführung einer Gewichtsfunktion in die obigen Relationen. Damit ergeben sich auch konsistente Schätzfunktionen für $C(\omega)$ und $\psi(\omega)$. Die Schätzungen wurden auch hier indirekt mit der oben angeführten Gewichtsfunktion von Tukey-Hanning durchgeführt.

Untersucht man z.B. mit Hilfe der Spektralanalyse die Kurse von 1000 Börsentagen¹⁾ einiger Aktiengesellschaften verschiedener Branchen, so zeigt es sich, daß alle geschätzten Spektren die gleiche Gestalt aufweisen: eine starke Konzentration spektraler Masse im Niederfrequenzbereich, flacher Verlauf der geschätzten Spektren im Mittel- und Hochfrequenzbereich. Die nachfolgende Graphik zeigt z.B. das geschätzte Spektrum der BASF.

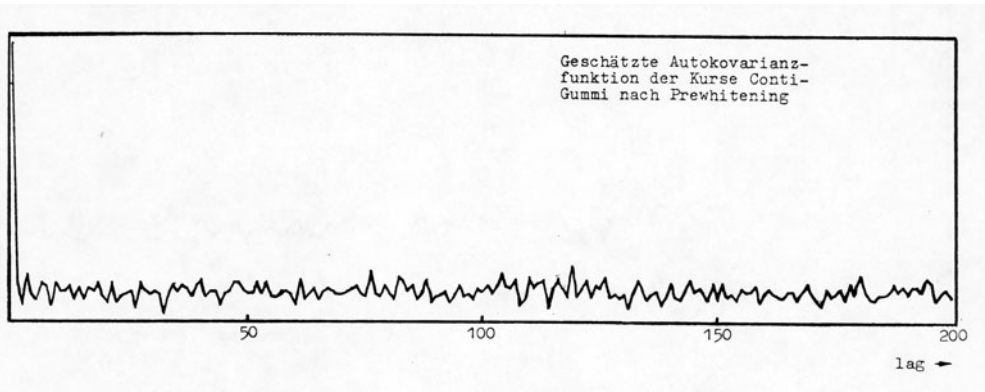


1) Es handelt sich dabei um die Kassakurse. Diese sind um Kapitalveränderungen etc. rückwärtsbereinigt.

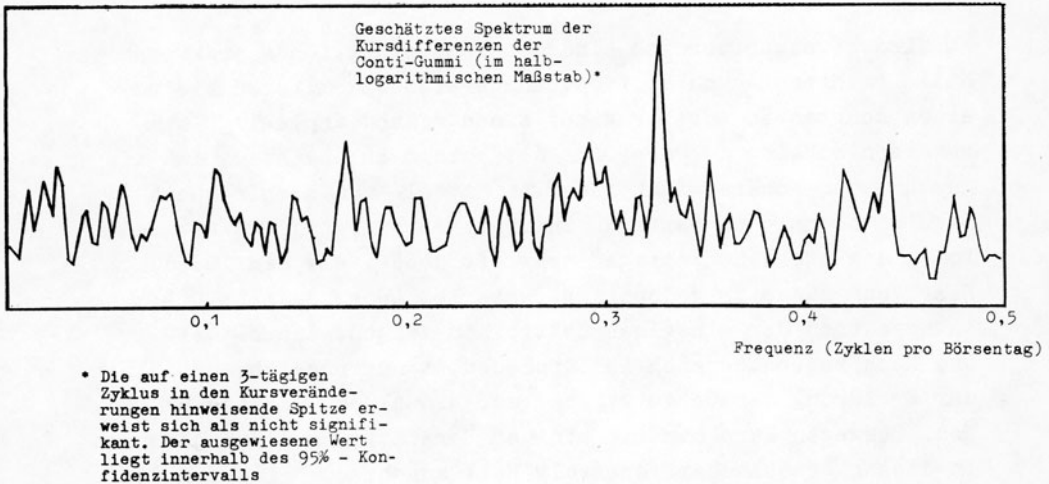
Da die vorliegende Zeitreihe nur eine Spanne von ungefähr vier Jahren umfaßt, soll zunächst auf eine Interpretation des Niederfrequenzbereichs verzichtet werden. Da sich in manchen Frequenzbändern das "leakage" stark bemerkbar machte, wurden die Reihen einer Transformation der Art

$$X_t - c X_{t-1}$$

unterworfen ("Prewhitening"). Die durch "Wiederfärben" ("Recolouring") gewonnenen Spektren unterscheiden sich aber prinzipiell nicht von den Spektren der Originaldaten. Bemerkenswert ist hier lediglich, daß die Autokovarianzfunktion der transformierten Reihen z.B. für $c = 0.99$ um Null oszillierte, wie aus folgender Graphik ersichtlich ist.



Diese Transformation erzeugt also schon praktisch einen Zufallsprozeß. Bildet man schließlich die ersten Differenzen der Kurse, so ergibt sich ein in allen Frequenzbändern flaches Spektrum, was nachstehende Graphik beispielhaft für die Kurse der Conti-Gummi zeigt:



Dasselbe Spektrum erhält man auch, wenn man die Daten zufällig mischt, denn dadurch wird ebenfalls die Trend-Zykluskomponente ausgeschaltet.

Aus den geschätzten Spektren läßt sich schließen: Kurzfristig ist die Kursentwicklung als reiner Zufallsprozeß anzusehen, der durch das random-walk-Modell beschrieben werden kann. Daneben dürfte noch eine Trend-Zykluskomponente von Bedeutung sein, was auch vom optischen Eindruck der Reihen nahegelegt wird. Somit können die Resultate von Granger/Morgenstern voll und ganz bestätigt werden.

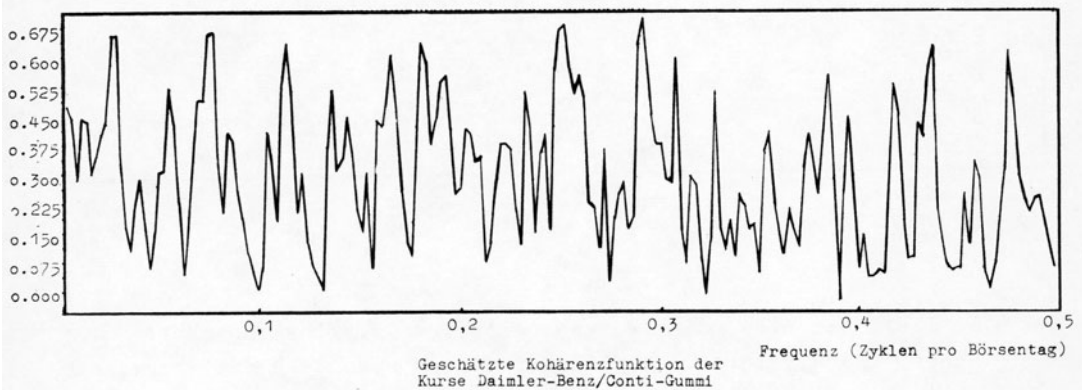
Beispielhaft wurden für die Kurse einiger Gesellschaften Kreuzspektren, Kohärenz- und Phasenfunktionen berechnet. Analog zum univariaten Fall bieten auch hier Kohärenz- und Phasenfunktionen für alle untersuchten Fälle das gleiche Bild:

Im Niederfrequenzbereich sind die Phasenfunktionen praktisch Null, im Mittel- und Hochfrequenzbereich oszillieren sie um einen konstanten Wert, was auf einen "fixed angle-lag" hinzuweisen scheint¹⁾. Dabei ist allerdings zu beachten, daß die geschätzten Kohärenzfunktionen im Mittel- und Hochfrequenzbereich nicht ohne weiteres interpretierbar sind, da sie infolge des starken "leakage" sehr oft größer als eins werden. Hier kann man sich jedoch auf zweierlei Weise helfen: Die Autospektren der einzelnen Zeitreihen zeigen, daß Mittel- und Hochfrequenzbereich im Durchschnitt nur ungefähr 4 - 5 % zur Erklärung der Gesamtvarianz der einzelnen Prozesse beitragen. Deswegen kann man mit einiger Berechtigung die Kohärenzen in diesen Frequenzbereichen als Null annehmen.²⁾ Eine andere Möglichkeit besteht darin, daß man die Kohärenzfunktionen aus den ersten Differenzen der vorliegenden Zeitreihen schätzt. Ein solcher linearer Filter läßt Kohärenz- sowie Phasenfunktionen unverändert, reduziert oder eliminiert jedoch das "leakage". Die so geschätzten Kohärenzfunktionen sind alle einheitlich interpretierbar: die Kohärenzen im Mittel- und Hoch-

1) Darunter wird ein konstantes Verhältnis zwischen Phasenverschiebung und Frequenz verstanden.

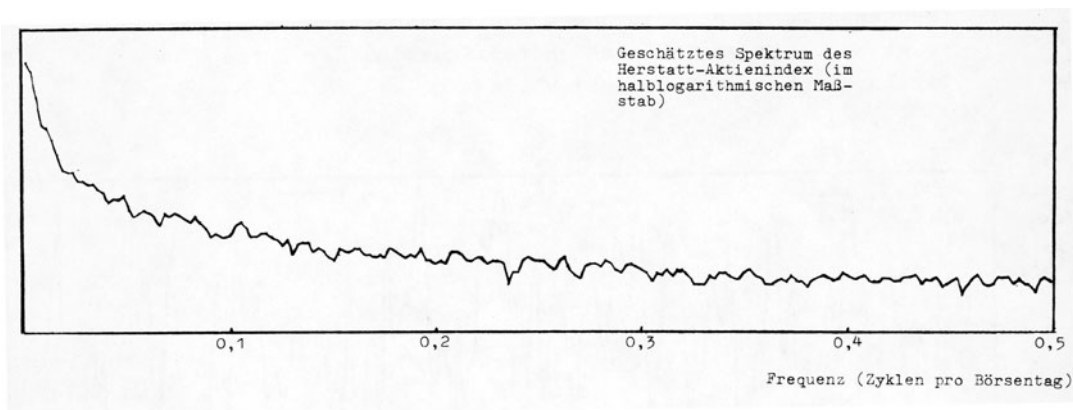
2) Eine derartige Interpretation empfehlen z.B. C.W.J. Granger und M. Hatanaka, *Spectral Analysis of Economic Time Series*, Princeton 1964, S. 102 -103

frequenzbereich sind durchweg klein. Nachstehende Graphik zeigt z.B. die geschätzte Kohärenzfunktion der Kurse Daimler-Benz / Conti-Gummi



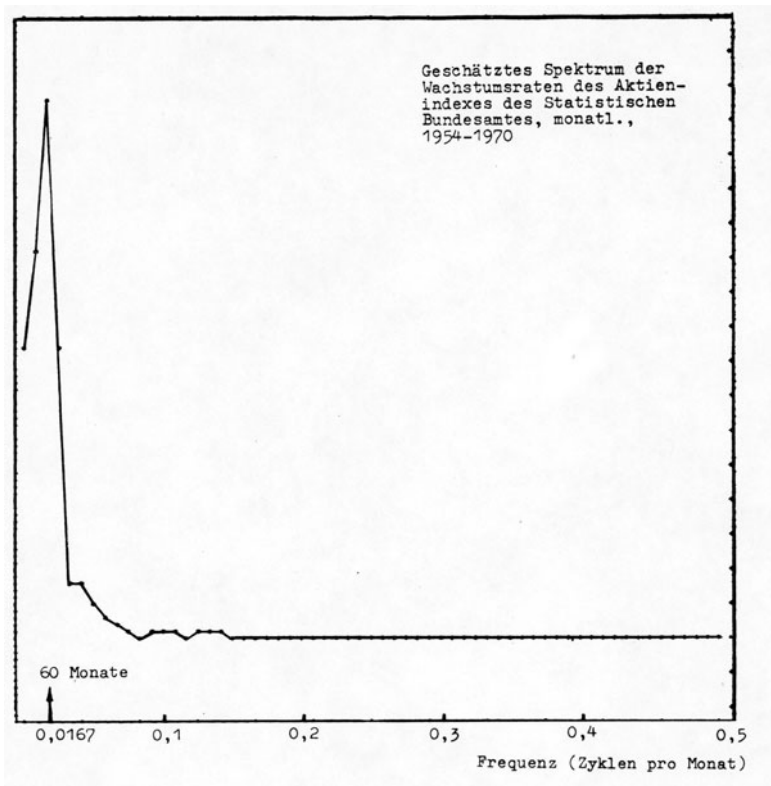
Somit ist ein „fixed-angle-lag“ nicht gesichert. Alle durchgeführten Kreuzspektralanalysen zeigen also, daß sich die Kurse der verschiedenen Papiere unabhängig voneinander entwickeln und daß insbesondere keinerlei gesicherte lead-lag-Beziehungen zu entdecken sind. Somit können die Ergebnisse von Granger/Morgenstern auch im Bereich der Kreuzspektralanalysen voll bestätigt werden.

Eine ganz andere Frage ist jedoch, ob die nun jetzt für Einzelpapiere gemachten Feststellungen auch für Aktienindizes zutreffen. Zur Untersuchung dieser Frage standen die Werte des Herstatt-Aktienindex für die gleiche Anzahl von Börsentagen zur Verfügung. Es zeigt sich auch dabei wiederum das gleiche Bild: eine starke Konzentration spektraler Masse im Niederfrequenzbereich, flacher Verlauf der geschätzten Spektraldichte im Mittel- und Hochfrequenzbereich.



Es liegt nun nahe, die festgestellte Trend/Zykluskomponente näher zu analysieren. Dabei interessiert besonders die Frage, ob sich der für die BRD festgestellte mittelfristige Konjunkturzyklus auch in den Börsenkursen wiederfindet. Leider standen dafür keine längeren Zeitreihen mit börsentäglichen Notierungen zur Verfügung. Deshalb wurde die weitere Analyse mit Hilfe der monatlichen Aktienindizes des Statistischen Bundesamtes und des Bankhauses Herstatt durchgeführt.

Da sich jetzt das Interesse in erster Linie darauf konzentriert, festzustellen, ob sich der bei anderen ökonomischen Zeitreihen für die BRD beobachtete Konjunkturzyklus von 60 Monaten auch in der Aktienkursbewegung nachweisen läßt, liegt es nahe, mit trendbereinigten Zeitreihen oder mit Wachstumsraten zu arbeiten. Es zeigt sich, daß sowohl die trenbereinigten Zeitreihen der genannten Aktienindizes, als auch die Zeitreihen der Wachstumsraten der Aktienindizes diesen Konjunkturzyklus von 60 Monaten ausweisen. Als Beispiel sei hier eines der geschätzten Spektren wiedergegeben.



Diese Spektren enthalten offensichtlich nur spektrale Masse in dem Frequenzband, das dem 60-Monatszyklus entspricht. Die Analyse zeigt also, daß sich die Kursentwicklung der untersuchten Papiere und die Entwicklung der einzelnen Aktienindizes durch ein Zeitreihenmodell der Form:

$$\text{Börsenkurs} = \text{Trend} + \text{Konjunkturzyklus} + \text{Zufallskomponente}$$

darstellen läßt. Nichts deutet auf die Existenz von Zyklen mit kürzerer Periodendauer als dem Konjunkturzyklus hin. Insbesondere ist kein saisonaler Einfluß festzustellen.

Investitionsentscheidungen

The Use of Simulation in the Evaluation of Alternative Designs and the Forecasting of Revenue for High Investment Transport Service Facilities

by A. M. Sutton, Epsom, U. K.

ABSTRACT

The Use of Simulation in the Evaluation of Alternative Designs and the Forecasting of Revenue for High Investment Transport Service Facilities

Modern methods of transport necessitate the use of highly complex service facilities, e. g. airport terminals, seaports, multi-storey car parks, etc. Invariably such complexes require substantial investment and as such should be well planned so that there will be a maximum return on capital employed. In order to design this type of facility a technique is required which can:

- (a) Determine the feasibility of the site and the building design.
- (b) Identify design faults and quantify the new effect of design changes.
- (c) Forecast the revenue in the long term obtainable from a specific design in a specific environment.

The technique of simulation has been developed in order to perform such a task. Examples are given of each of these applications and a consideration is given of the future uses of simulation in the field of iterative design.

1. Introduction

The size of modern transport systems and the high demand placed upon them, causes the associated service facilities (e. g. Airport Terminals, Seaports, Multi-storey Car Parks, etc.) to be highly complex. This complexity, together with the frequency of new design requirements, causes the conventional design systems based upon previous experience to lose much of their power, since prior experience can only assist the designer of a new type of facility to a limited extent. The W. S. Atkins Group has been conscious of this problem for some time and has developed tools in order to assist the designer to solve this problem.

The designer of a service facility requires a technique which can:

1. Test site feasibility.
2. Test design feasibility for present and future requirements.
3. Identify design faults.
4. Quantify the effect of design changes.
5. Forecast facility revenue.

Several different techniques may be applied in order to obtain a solution of these problems, i. e. real life modelling, data adaption, manipulating micro models, using queueing models, simulation, etc.

A real life modelling approach has been used by some consultants in order to investigate container handling systems for the Channel Tunnel terminal. This technique consists of undertaking experiments with equipment similar to that which may finally be used. Such a form of modelling has the advantage that less validation is required, since only the working environment is being simulated.

However, the cost of modelling several systems and determining the effect of design changes is obviously prohibitive in most cases.

Data adaption is a technique in which existing data is altered in such a way as to mirror the changes that are required to reach the new situation. Thus, entrance and exit flow figures for a small multi-storey car park may be modified in order to attempt to investigate the required number of entrance and exit channels in a larger car park. The obvious advantage of such an approach is the low cost involved: however, the difficulties of correct data adaption are numerous, i. e. complex interactions may be overlooked.

A further technique which can be adopted when modelling complex service facilities is the manipulation of micro models to form a simple macro model. Thus in modelling, an airport terminal, for example, individual simulation models may be constructed of specific facilities such as immigration area, customs, etc. When design changes are proposed in one facility these are incorporated in the associated model, and the output of this model is fed into the other models in order to determine the effect that design change has upon the total facility. The main advantage of this type of technique is the ease of modelling a complex system by splitting it into modules. However, such an approach often takes little account of the strong interaction which can occur between individual model modules, and can lead the user to the view that a set of optimally designed individual facilities leads to an optimum design of the total facility.

Queueing theory may be applied to these problems and has the major advantage that the output from a queueing model is a statistic and thus many runs of the model are not necessary in order to obtain a definition of the output distributions. However such a technique has difficulty in modelling the interference which can occur in many service facilities; and can only be used in a limited number of applications.

Simulation is a technique which avoids the disadvantages discussed above, and thus it was the approach which the W. S. Atkins Group has adopted to produce models of service facilities. It has been developed into an iterative design tool which can quantify design changes, enabling the designer to identify the effect of these changes and thus to experiment with the overall design by repeating the

process after obtaining quantitative results. This form of technique enables the designer to readily identify potential problem areas and to try various forms of solution.

The technique of simulation has the principal advantage that it may be used to model situations in which complex interactions arise. In addition, simulation is an approach which can be readily understood by non-mathematicians and consequently accepted by them. It thus fulfils the prerequisite of any design tool in that the designer accepts the usefulness of the technique and is prepared to test and use it. The major difficulty encountered when using simulation in this field is that of validation. This problem should never be underestimated, much ingenuity and possibly a little faith is required when the form of facility being modelled is entirely new.

In order to establish an acceptable and logically correct model, two techniques have been found to be successful. The first is the use of hand simulation models and associated diagrammatic representations to demonstrate the principles and logic of the proposed system. Such an approach was adopted when modelling a multi-storey car park and it was found to assist the designer greatly in understanding the problems resulting from a specific design. Secondly, transportation and other consultants have been involved throughout the model-building stage, in order that simple, meaningful and acceptable models can be built. More detailed models may later be constructed if there is necessity. However, it is often found that such sophisticated models shed very little further light on the problem.

Three examples will be used in order to demonstrate the use of simulation in design process: the first concerns the use of simulation in determining the feasibility of a site for the English Channel Tunnel Terminal and to determine a feasible arrangement of the service facilities. The second example is the evaluation of alternative designs of a multi-storey car park and shows how the technique of simulation may be used in an iterative fashion to seek a best design. It should be noted that the word 'best' does not denote the optimum, merely a 'best' of those considered when judged by the specific set of criteria chosen by the designer. Lastly, the possibility of using simulation as a technique to assist in the forecasting of revenue from the facility is briefly considered.

2. Testing site and design feasibility

In 1968 the W. S. Atkins Group was invited by the Ministry of Transport to undertake a study to consider the location of the British Terminal of the Channel Tunnel.

Two sites were originally considered by the Ministry of Transport. After initial investigation had been carried out, it was decided that one of the sites was preferred from the operating point of view. However, there was a major difficulty, that the site was subject to several physical restraints. Thus it was necessary to determine if it was feasible to place the required facilities for such a terminal within the site area, and also to determine the optimum arrangements within this area in order to satisfy service demand.

The layout of the site itself required full consideration of the internal circulatory systems, which account for a large proportion of the area. Final geometric layouts of the site for the expected demand and proposed service processes were prepared for each of three possible layouts. The difference between these layouts depended on whether immigration, customs and health facilities were situated at the French terminus only, at the British terminus only, or alternatively at both.

A simulation⁴ model was developed of each layout in order to determine the delays likely to be encountered at each facility due to variation in service time and flows. Two total demand figures were used when the model was being run, the first being the forecast for the opening date of the tunnel; the second a forecast for the distant future.

The modelling process commenced with a consideration of the sequence of facilities required for each of the three possible arrangements of customs, immigration and health control. These sequences were determined by reference both to the data collection exercise and also discussions with interested parties. The analysis continued by matching the assumed demand for each facility and the modal split, having determined the facility performance during data collection. (The modal split is the proportion of each vehicle and passenger type within the total demand.) It was found at this stage that the land area demand was highly sensitive to the requirements of commercial vehicles, thus

sensitivity tests were used in this area of the model with three levels of commercial vehicle activity. The input of these demands to the model determined the number of channels required within each facility and the consequent queues in each facility. The total waiting time of passengers in the system determines the feasibility of this number of channels and, if necessary, further iterations of the process were carried out. The total number of channels determined the ground area and staff requirements.

With the known facility sequence, queue requirements and design of each facility determined, a layout of the site was prepared for each of the three possible customs, immigration and health arrangements. This was done while taking into account the site constraint, i.e. considering the requirements of site roads, access roads and railway tracks. A flow diagram of the modelling process is shown in Figure 1.

The result of this study was a feasible layout within the site, with, in addition, recommendations concerning the extra requirements that the facility placed upon the site as the demand changed in the future.

There are two aspects of this study which are of note. Firstly, the iterative nature of such a project, in which all interested parties were consulted and concerned with all stages of the model-building, experimentation, and analysis of results. It is felt that results, presented in a suitable form, are extremely valuable in generating feasible and possibly optimising design suggestions. Secondly, as a result of undertaking this study, of using the model, and of discussing results with designers, other possible methods of fulfilling the demand were suggested. These ideas were evaluated and produced in a further part of the report.

In this study the part played by the actual simulation model may be considered small. However, the design model was produced in the light of Operational Research attitudes and it is this concept, together with the idea of an iterative design model, that should be emphasised.

3. The evaluation of alternative designs

In 1970 the W. S. Atkins Group was asked by the Harlow Development Corporation to evaluate several alternative designs of a multi-storey car park.

The site and size of the proposed car park had been established and the principal objective of the study was to evolve a technique for estimating in numerical terms the differences between the alternative car park layout designs that had been submitted.

It is difficult to analyse this problem using conventional mathematics because the times between arrivals of vehicles at a car park and the times associated with their individual travelling and parking manoeuvres are variable, and consequently, queueing and other vehicle interaction occurs within the car park. However, the problem is amenable to solution by simulation and there are three aspects of the problem which are sufficiently independent to be studied separately.

1. The internal layout design
2. The number of configurations
3. The number, capacity and type of lifts installed

In order to ensure that the simulation model correctly represents the operation of the car park, two distinct types of data are required. First, those that are peculiar to the town or parking area within which the car park will be sited, e.g. arrival pattern of vehicles at the garage, vehicle occupancy distribution and the distribution of the length of stay time.

The second category of data is that which is related to the usage of multi-storey car parks in general, i.e. time taken to park and manoeuvre, percentage of vehicles reversing into parking bays, speed of vehicles along unobstructed aisles, up and down ramps, and round corners.

In the simulation each of the movements that cars perform during normal usage of the car parks is modelled. To illustrate the technique a car moving in a car park with completely unobstructed aisles may be considered. The car is assumed to travel along aisles, round corners and up ramps at the mean observed speeds until a vacant bay is located. The bay chosen by the driver need not be the first encountered; it may be such as to minimise walking distance to a pedestrian exit. At the chosen bay the car will commence obstructing the aisles for the obstruct time appropriate to the constraints encountered (i.e.

vehicles already parked or alternatively pillars). When the car enters the bay after completing its period of obstruct parking, the non-obstruct parking time starts. The end of this period is signified by the departure of the occupants from the vehicle. At this point the occupants are assumed to walk to the lift on their floor at the mean observed speed while the vehicle starts its length of stay time. Events at the end of the stay time when the occupants return to the vehicle, follow the converse of the arrival sequence.

Initially two design layouts were simulated for the Harlow Development Corporation - both designs had a total capacity of approximately 1,200 spaces. The first of these (Figure 2) is a split-level design in which vehicles entering the car park pass all available bays on their journey to a vacant bay: thus there is no need for a control system. The second layout (Figure 3) is a continuous ramp design which, due to siting constraints, lacked an express down ramp.

The results obtained from these layouts are shown in the table. The 'time to park' is defined to be the time which elapses between the vehicle entering the car park and finally reaching its parking bay. The 'time to leave' is defined to be the interval between the departure of the vehicle from the parking bay and its arrival at the car park exit. The 'total travel time' is the sum of these two values.

Using the parameters shown in the table to judge the design, it can be seen that the split-level layout is considerably better than the continuous ramp design. Similarly, it can be seen that with such a large car park this continuous ramp design may only become acceptable if a fast exit ramp is added. It was thought that greater improvements would result from the adoption of fast entrance ramps together with an appropriate control system for directing vehicles. Therefore, a third design (Figure 4) was produced, which contained a single fast up ramp together with two express down ramps. It can be seen from the results that the adoption of such a design produces, as would be intuitively expected, great time savings for vehicles entering the car park, and rather smaller improvements for vehicles leaving. From a study of these three designs, it appears that the following design objectives should be incorporated in a good design:

- (a) Vehicles travelling to park should pass a minimum number of bays;
and
- (b) There should be a minimal possible vehicle interference.

Consideration of these criteria led to the production of a further design (Figure 5) in which the bays were orientated such that there was minimum interaction between vehicles travelling within the car park and those parking. The table shows clearly the time savings that may be obtained by vehicles using this design. Thus it can be seen how the technique of simulation may be used by the designer to evolve a layout design which is suitable for the environment in which the car park will be located.

4. Revenue forecasting

Often the service facilities described in this paper are not required to make a profit. However, it is frequently necessary to forecast the revenue resulting from such a facility in order to justify further expenditure which improves the service. Also, if charges are to be made for the use of this facility, it is obviously necessary to determine the optimum charging policy.

The simulation model of a multi-storey car park described above may be used in order to forecast the revenue obtainable from the facility. However, such a model may not be directly used since different pricing structures may affect demand. In fact a specific pricing policy may be chosen because it is thought that it will, for example, discourage long-term car parking. In order to determine the variation of demand with price, some form of market survey may be used. The recent work published by Heald* suggests that an approach using multiple regression techniques is applicable, and would enable the effect of the local siting of shops belonging to major merchandising chains, upon a car parking demand price curve, to be determined. This is a major factor since often such shops are willing to share in the finance of car parking if it can be proven that their own trade will benefit. Since the simulation model outputs the total time that vehicles spend in the car park, the addition of a car parking price module will enable the revenue to be determined when using a specific pricing policy.

5. Conclusions

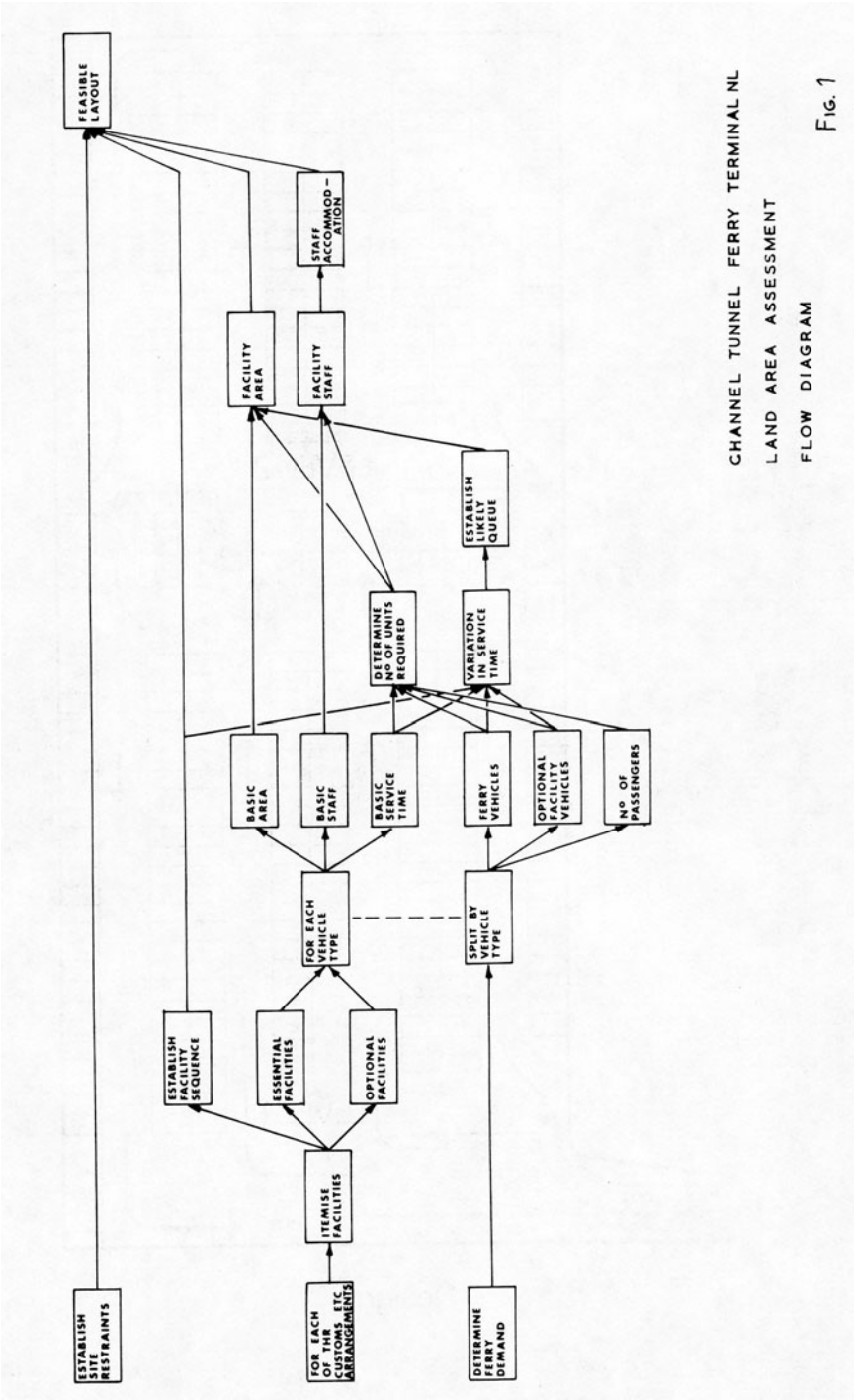
Designers of service facilities have found that the technique of simulation has been of assistance in ensuring that the facility functions efficiently throughout its working life. Such service facilities are becoming larger and more complex and thus require a formal design tool in order that a satisfactory design may be obtained quickly and cheaply. This tool may be used most effectively in an iterative manner.

The iterative design process is one in which there is a dialogue between the model and the designer in order that the designer may readily determine the effect of specific design changes. The type of models discussed above are only in their infancy, since at present they are in no sense formal and automatic. However, one might envisage rather more automated design systems in which location parameters depending on the local environment, together with local data, are used as input to a generalised model of, for example, an airport terminal. Such a model would be used to investigate the overall feasibility of a specific facility on that site.

The technique of simulation when applied to a design system has four important contributions which assist the planner:

1. It enables him to make quantitative comparisons between alternative layouts using distribution which refers specifically to the environment in which the facility will be built.
2. He may identify design faults through the diagrammatic representations of the layout which may be produced by the system.
3. Modifications of inputs to the system may be used to investigate the likely effects of variations in future demand.
4. Such a system may be used to determine the optimum pricing policies in order to forecast the likely revenue generated by the facility.

* G. Heald 'The Application of A.I.D. and Multiple Regression Techniques to the Assessment of Store Performance and Site Selection', British O.R. Conference, 1971.



CHANNEL TUNNEL FERRY TERMINAL NL
LAND AREA ASSESSMENT
FLOW DIAGRAM

Fig. 1

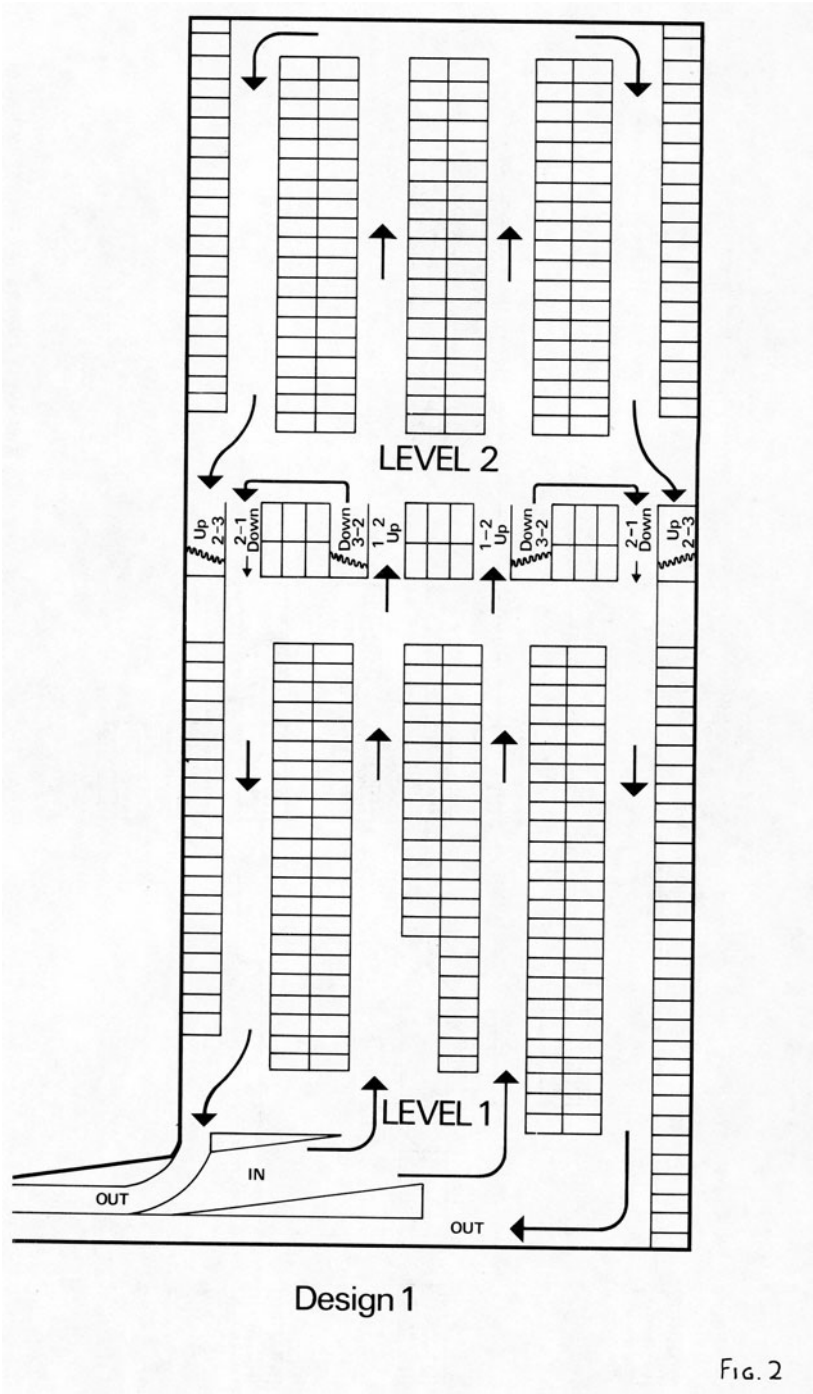
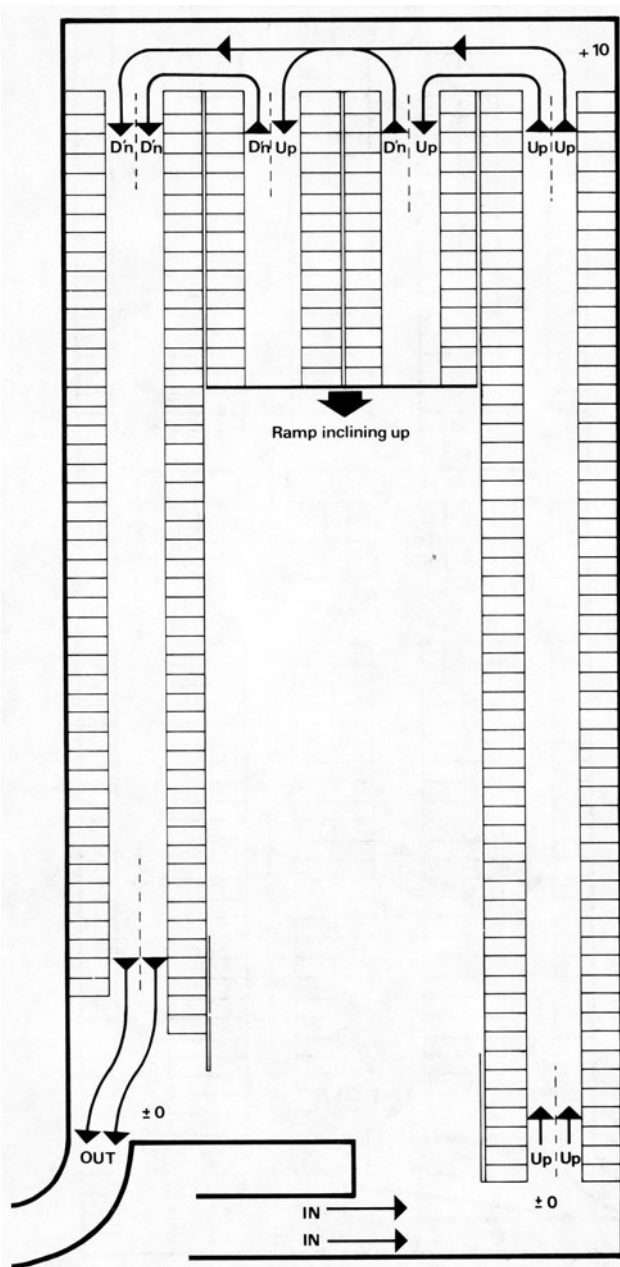


FIG. 2



Design 2

FIG. 3

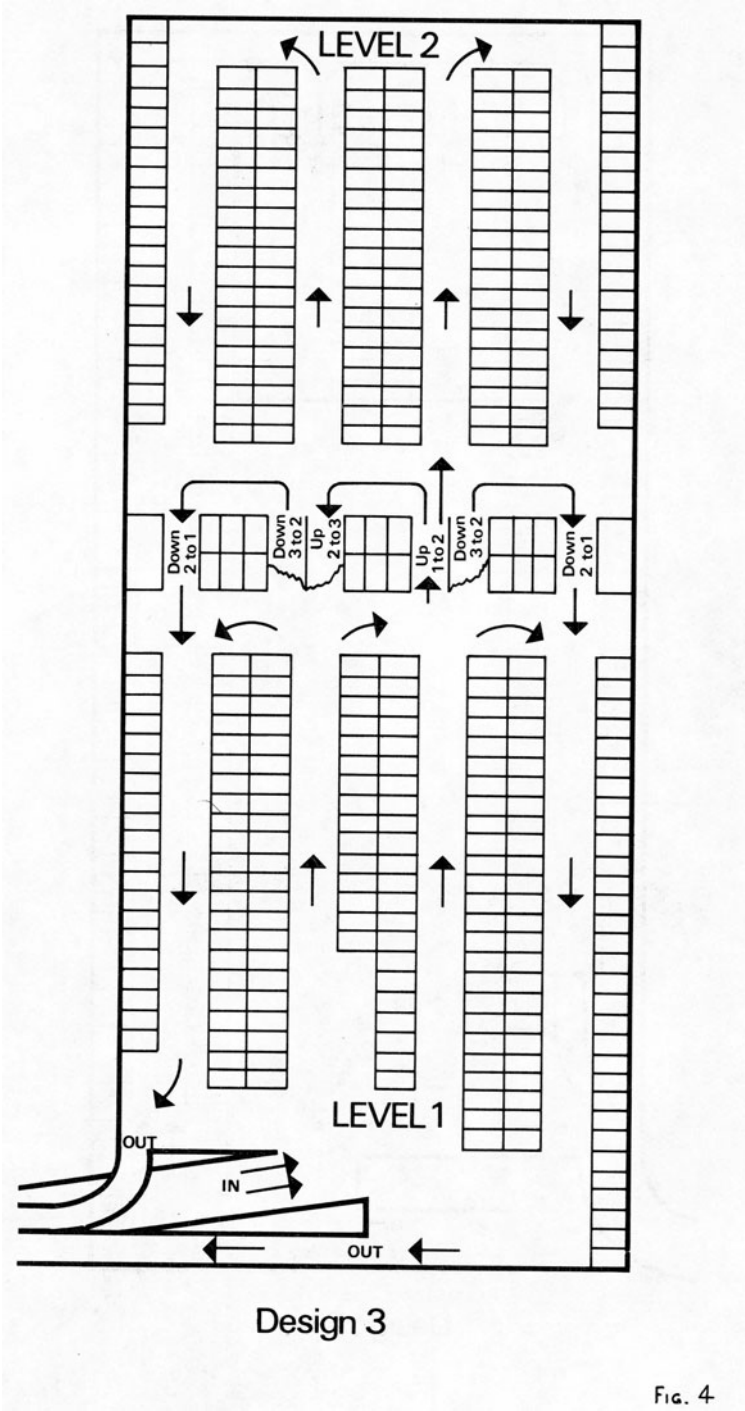


FIG. 4

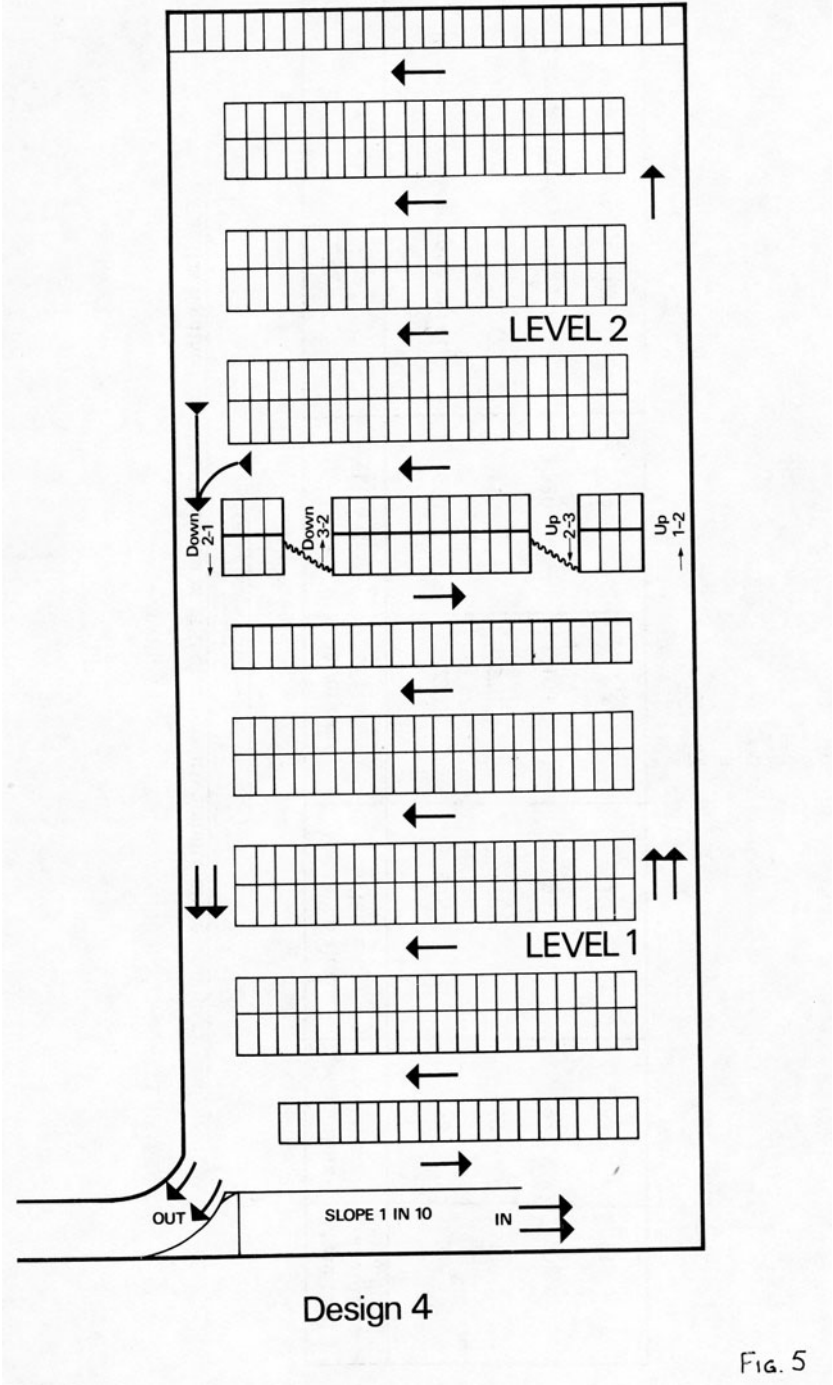


FIG. 5

TABLE 1.
Results obtained from the simulation of four alternative multi-storey car park designs
(all times expressed in sec.)

	Maximum parking time	Mean time to park	Maximum time to leave	Mean time to leave	Mean total travel time
Split-level - no fast up-ramp (1)	590	230	475	95	325
Continuous ramp (2)	510	270	410	160	430
Split-level - three fast ramps(3)	330	106	335	98	204
Split-level - two fast ramps (4)	175	85	110	53	138

Probleme der Fertigung

Heuristische Methoden zur Arbeitsangleichung bei Fließbändern
 von L. Kosten, Delft

1. Das Problem der Arbeitszeitangleichung bei Fließbändern

Bei einem Fließband ist es wichtig, die gesamte Arbeitszeit, die für die Anfertigung eines Werkstückes benötigt ist, möglichst gleichmäßig auf die in Reihe geschalteten Arbeitsplätze zu verteilen. Denn der Arbeitsplatz, wo pro Werkstück am längsten gearbeitet wird, hält die anderen Plätze auf und bestimmt die geringste Durchschubzeit und somit das Arbeitstempo.

Die Gesamtarbeit pro Werkstück kann aufgeteilt werden in eine Zahl von Arbeitselementen U_1, \dots, U_m , deren Einzeldauern t_1, \dots, t_m bekannt sind. Die t_i werden als ganzzahlig (etwa in Minuten gemessen) vorausgesetzt. Technologisch gibt es gewisse Bedingungen für die Reihenfolge, in welcher die Arbeitselemente ausgeführt werden (Ordnungsrelationen). Als Beispiel wird das (einer Arbeit von Tonge [1960] entnommene) Beispiel der Figur 1 gewählt mit 21 Elementen, welche durch 6 Arbeitsplätze zu bedienen sind. Die t_i sind die über den Kreisen geschriebenen Zahlen. Die Pfeile stellen die technologisch vorgegebenen Ordnungsrelationen dar (z.B.: U_{21} kann nicht anfangen bevor U_2 und U_4 beide fertig sind).

Es seien L_1, \dots, L_n die Arbeitsplätze (numeriert vom Anfang bis zum Ende des Fließbandes). Jedes Arbeitselement soll jetzt einem Arbeitsplatz zugeordnet werden. Hierbei ist jedoch damit zu rechnen, daß ein Element U_i nur dann einem Platz L_j zugeordnet werden kann, wenn alle Vorgänger von U_i im Relationsdiagramm je einem der Plätze L_1, \dots, L_j zugeordnet werden. Eine vollständige Zuordnung der Elemente wird realisierbar genannt, wenn die gegebenen Bedingungen eingehalten sind. Jedem Arbeitsplatz L_j fügen wir eine Zahl S_j zu, die Summe aller t_i , welche an L_j zugeordneten U_i entsprechen. Eine realisierbare Zuordnung ist optimal, wenn die größte der S_j ($j = 1, \dots, n$) minimal ist (Minimaxprinzip). Die zu minimalisierende Größe $\text{df} \text{ Max } S_j$ wird weiterhin Zielfunktion genannt. Im folgenden

werden wir jeden Arbeitsplatz identifizieren mit der Menge der ihm zugeordneten Elemente.

Gemäß der obigen Beschreibung ist im Beispiel der Figur 1 die folgende Zuordnung:

$\{L_1 = U_1, U_3, U_4\}$	$S_1 = 18$
$\{L_2 = U_5, U_7\}$	$S_2 = 17$
$\{L_3 = U_6, U_8, U_9, U_{11}\}$	$S_3 = 19$
$\{L_4 = U_{10}, U_{12}, U_{15}, U_{13}, U_{16}, U_2\}$	$S_4 = 18$
$\{L_5 = U_{17}, U_{21}\}$	$S_5 = 20$
$\{L_6 = U_{18}, U_{14}, U_{20}, U_{19}\}$	$S_6 = 13$

realisierbar mit $F = 20$ (siehe auch Figur 2). Die Vermutung liegt jedoch nahe, daß die Angleichung noch nicht optimal sei.

Es gibt manchmal noch zusätzliche Bedingungen folgender Art. Irgendwo am Fließband muß ein automatischer Arbeitsvorgang stattfinden. Dieser braucht wohl Zeit, jedoch keine menschliche Hilfe. Es ist dann angebracht, diese Station zwischen zwei Arbeitsplätzen anzuordnen und ihr eine Durchschubzeit gleich einer Durchschubzeit der Arbeitsplätze zuzuteilen. Die Strecken am Fließband vor und nach dieser automatischen Station werden wir die Zonen (A und B) nennen. Von vornherein braucht noch nicht festzustehen, wieviel Arbeitsplätze beide Zonen umfassen. Nun kann aus technischen Gründen festliegen, daß bestimmte Elemente nur zugeordnet werden können an Plätze, die der Zone A (oder B) zugehören. Für andere Arbeitselemente können beide Zonen zuständig sein. Diese Zonierung kann sich auf mehr als zwei Zonen ausdehnen (die jedoch alle eine ganze Zahl von Arbeitsplätzen umfassen). Für jedes Element sind dann die zulässigen Zonen spezifiziert. In komplizierten Fällen brauchen die Arbeitsplätze einer Zone nicht anschließend zu sein. Als Beispiel wählen wir die Zonierungsindizierung von Tonge für den Fall der Figur 1:

Element:	1	2	3	4	5	6	7	8	9	10	11	12
Zulässige Zone:	A	A	AB	AB	A	B	AB	AB	B	B	B	B
	13	14	15	16	17	18	19	20	21			
	BC	C	BC	BC	C	BC	C	C	A			

Zu den Realisierungsbedingungen kommt dann noch die Vorschrift, daß alle einem Arbeitsplatz zuzuordnenden Elemente einen Zonierungsbuchstaben gemein haben. In Figur 3 ist eine realisierbare Zuordnung für diesen Fall gegeben mit $F = 20$.

Das zu lösende Problem kann jetzt allgemein folgendermaßen beschrieben werden:

- Wie findet man eine realisierbare Zuordnung?
- Wie findet man eine optimale Zuordnung, d.h. eine realisierbare Zuordnung mit minimalem Wert der Zielfunktion F ?

Die Lösung der Frage b) kann zurückgeführt werden auf Frage a). Man sucht eine realisierbare Zuordnung unter der zusätzlichen Bedingung, daß kein S_j größer sei als eine gewisse obere Schranke B_u (wenn B_u groß ist, ist dies keine reelle Begrenzung). Wenn eine realisierbare Zuordnung gefunden ist, verringert man B_u um eine Einheit und wiederholt den Suchprozeß. Wenn der letztere scheitert, enthält die vorige Stufe eine Optimalzuordnung.

Wir werden in den nächsten Abschnitten immer wieder Figur 1 als Schulbeispiel benutzen. Es braucht kaum gesagt zu werden, daß in der Praxis die Dimensionen größer sind (einige Hunderte von Elementen, ein Paar Zehner von Arbeitsplätzen und bis zu 10 Zonen).

2. Mögliche Lösungsmethoden

Allererst kann man fragen nach einer einwandfreien mathematischen Formulierung. Diese könnte etwa die folgende sein:

Gegeben sei:

- a) eine Elementenmenge $U = \{U_1, \dots, U_m\}$;
- b) den Elementen zugeordnete natürliche Zahlen: $U_i \rightarrow t_i$;
- c) eine Anzahl Ordnungsrelationen: $U_{i_k} < U_{j_k} \quad (k = 1, \dots, p)$;
- d) eine Anzahl Teilmengen $Z_i \subset U \quad (i = 1, \dots, q)$ mit

$$\bigcup_{i=1}^q Z_i = U;$$

Es sei zu bestimmen:

eine geordnete Zerlegung L_1, \dots, L_n von U , d.h.:

- e) $U = \bigcup_{i=1}^n L_i$;
- f) $i \neq j \Rightarrow L_i \cap L_j = \emptyset$;
- derart daß:
- g) $U_i < U_j \wedge U_j \in L_k \Rightarrow U_i \in \bigcup_{p=1}^k L_p$;
- h) $(\exists i): L_k \subset Z_i \quad (k = 1, \dots, n, ,$
- j) Wenn $S_j \stackrel{\text{df}}{=} \sum_{i \mid U_i \in L_j} t_i$, so ist $\max_{(j=1, \dots, n)} S_j$ minimal.

Die Beschreibung kann noch beliebig weiter sophistiziert werden, indem man jeglichen klaren Text ausmerzt. Man bekommt dann ein Musterbeispiel mathematischer Obstruktion, das einerseits das Problem des Ingenieurs völlig verschleiert, andererseits keinen Anhaltspunkt zu bieten hat für eine praktische Lösung. Es ist ebenso "exakt" und deshalb viel vernünftiger, das Problem und dessen Lösung in der Originalfassung zu besprechen.

Das vorliegende Problem gehört zu den kombinatorischen Problemen der Unternehmensforschung. Die klassische Mathematik bietet hauptsächlich "existentielle" Kombinatorik, d.h. sie befaßt sich mit Fragen nach der Zahl der Lösungen für bestimmte Probleme. Die für die Praxis meist anwendbare Methodiken für kombinatorische Probleme der Unternehmensforschung sind:

1. Erschöpfung (exhaustion): Ausprobieren aller Möglichkeiten; meistens nicht brauchbar wegen der großen Zahl dieser Möglichkeiten.
2. Direkte analytische Lösung: Analyse liefert eine direkte Regel zur Bildung der Lösung; Beispiel: Johnsons Regel für das Zwei-Maschinen Problem.
3. Verbesserungsalgorithmen: Beispiel: die "Hungarische Methode" für das einfache Zuordnungsproblem.
4. Branch-and-Bound Methoden: Dies ist eigentlich die Methode der Erschöpfung, wobei jedoch durch eine geeignete Suchstrategie die Zahl der zu probierenden Fälle meistens drastisch verringert wird; Beispiel: die Methode von Little u.a. für das Problem des Handelsreisenden.
5. Heuristisches Verfahren:
6. Monte-Carlo Methoden:
7. Lernverfahren:

Die ersten vier Methodiken sind als "analytisch" zu bezeichnen in dem Sinne, daß theoretisch das Finden einer Lösung gewährleistet ist, falls eine Lösung besteht. Die letzten drei kommen erst in Frage, wenn die ersten vier praktisch versagen. Für heuristische und Monte-Carlo Verfahren gilt: "manchmal geht's, manchmal geht's nicht". Man muß sich dann damit begnügen, daß z.B. in 90% aller Fälle eine Lösung gefunden wird, oder daß eine "nahezu" optimale Lösung gefunden wird.

Da der Verfasser der Absicht war, daß das Angleichungsproblem der Fließbänder zu widerspenstig sei (allerdings bei Problemen normaler Größe) um analytische Methoden zuzulassen, hat er sich gleich um heuristische und Monte-Carlistische Behandlungsweisen bemüht^x. Teilweise sind diese dem Schrifttum entnommen.

^x) Siehe jedoch R.W. Sierenberg.

3. Tonges Methode

Die Methode von Tonge wird kurz erörtert werden für den Fall ohne Zonierung. Als Ziel wird gesetzt werden, eine Zuordnung zu finden mit $B_u = 20$ (der mittlere Wert der S_j ist $105/6 = 17 \frac{1}{2}$). Tonge hat zuerst eine Art Kodifizierung des Relationsdiagramms eingeführt. Dies wird beschrieben durch eine Hierarchie von Reihenschaltungen, Parallelschaltungen und sog. "Z-Schaltungen", deren feinste Komponenten die Arbeitselemente sind. Eine Z-Schaltung ist ein Gebilde mit zwei vorderen Komponenten und zwei hinteren Komponenten. Eine der hinteren Komponenten ist Nachfolger beider vorderen Komponenten, die andere hintere Komponente folgt nur auf einer der vorderen Komponenten. Mit Hilfe dieser Strukturelemente kann Figur 1 etwa umgeformt werden in Figur 4, wobei jedoch die Ordnungsrelationen $U_{13} < U_{17}$ und $U_{14} < U_{19}$ vernachlässigt werden müssen (in der Hoffnung, daß die zu findende Zuordnung diesen Relationen "von selbst" genügen wird).

Nun verteilt Tonge die 6 Arbeitsplätze möglichst gut auf die Hauptkomponenten (siehe Figur 4). Dies weist einen Arbeitsplatz mit 4 Minuten und 5 Plätze mit 101 Minuten auf. Nun können die vordere Komponenten einer Z-Schaltung frei in der Hierarchie "nach vorne" verschoben werden, wobei die neuen Gebilde wieder mit den kodifizierten Strukturelementen beschrieben werden können. Dadurch kann Figur 4 umgewandelt werden in Figur 5. Diese besteht aus einer Reihenschaltung von 4 Komponenten: U_1 , U_3 , U_4 und dem "Rest". Auf die Menge der ersten drei Komponenten entfällt eine Arbeitszeit von 18 Minuten, was ausreicht. Somit hat man das "kleinere" Problem um $105 - 18 = 87$ Minuten auf "den Rest" zu verteilen. Auch für Parallelschaltungen lassen sich Verschiebungsverfahren angeben. Durch wiederholte Umbildung der Struktur mit Hilfe dieser Verfahren läßt sich schließlich eine Zuordnung mit $F = 20$ finden (welche nicht dieselbe ist wie Figur 2). Durch weitere Austauschvorschriften läßt sich die gefundene Lösung noch verbessern.

Der Verfasser hat keine Erfahrung mit dem Tongeschen Verfahren, zumal auch weil Teile des Verfahrens nur skizzenhaft gege-

ben worden sind. Seines Erachtens ist das zunächst zu beschreibende Verfahren einfacher und viel mehr zielgerecht.

4. Die "Positional Weight" Methode

Von Helgeson und Birnie ist eine Methode beschrieben worden, die mit "Positional Weights" (PW) arbeitet. Der Positional Weight PW_i eines Arbeitselements U_i wird definiert als die Summe von t_i und die t_j aller Elemente U_j , die im Relationsdiagramm entweder mittelbar oder unmittelbar auf U_i folgen. So ist z.B.:

$$PW_{13} = t_{13} + t_{17} + t_{20} + t_{18} + t_{19} = 28.$$

In Figur 6 sind die PW's aller Elemente unter den Kreisen schräg angeschrieben. Jetzt werden eine untere Schranke B_L (sag 17) und eine obere Schranke B_u (sag 20) gewählt. Nun werden die Arbeitsplätze L_1, L_2, \dots, L_6 der Reihe nach in dynamischer Weise mit Arbeitselementen belegt. Solange ein Arbeitsplatz noch weniger als $B_L = 17$ Minuten Arbeitszeit enthält, werden noch mehr Elemente hinzugefügt, wenn möglich.

Eine Zufügung ist möglich, wenn es die Realisierungsbedingungen zulassen und die Gesamtzeit für den Arbeitsplatz durch die Zufügung nicht größer als $B_u = 20$ Minuten wird. Falls mehr als ein Element kandidiert für die Zufügung, wird dasjenige gewählt mit dem größten P.W. In dieser Weise wird für den Fall ohne Zonierung die in Abschnitt 1 gegebene realisierbare Zuordnung gefunden. Dort sind für jeden Arbeitsplatz die Elemente gegeben in der Reihenfolge, in welcher sie kraft der gegebenen Vorschriften hinzugefügt werden müssen. Wenn z.B. L_1, L_2 und L_3 gebildet worden sind und L_4 schon U_{10} und U_{12} enthält, wird U_{15} zugefügt, weil $PW_{15} = 31$ größer ist als die PW's der übrigen Bewerber: $PW_2 = 10$, $PW_{13} = 28$ und $PW_{14} = 5$. Nach diesem Erfolg wird das Verfahren versucht mit $B_L = 17$, $B_u = 19$. Auch dies war möglich. Insgesamt kostete das 10 Minuten Arbeit ohne Hilfsmittel. Ein (auch der Tongeschen Arbeit entnommenes) Beispiel mit 70 Elementen und 22 Arbeitsplätzen wurde ohne Zonenbedin-

gungen mit Hilfe eines Tischrechners in etwa 2 Stunden gelöst.

Jedoch kann das Verfahren versagen auch in Fällen, wo eine Lösung möglich ist. So wurden für den Fall ohne Zonierung für $B_z = 17$, $B_u = 18$ als auch für den Fall mit Zonierung für $B_z = 17$, $B_u = 20$ keine Lösungen erhalten, obwohl sie bestehen (siehe den nächsten Abschnitt und Figur 3).

6. Nochmals die P.W. Methode; Erweiterung

Elemente mit großem P.W. haben viele Nachfolger. Nun steckt die "hohe Weisheit" der P.W. Philosophie hierin, daß "durch die Wahl von Elementen mit großem P.W. während der dynamischen Zuordnung möglichst viele Türen in die Zukunft geöffnet werden". Heuristische Methoden beruhen immer auf derartigen einleuchtenden Überlegungen. Doch eben dieses Bild der "sich öffnenden Türen in die Zukunft" ist der Schlüssel zu Erweiterungen.

Die P.W. Methode angewandt auf Figur 1 (ohne Zonierung) mit den Schranken 17/18 versagt mit folgender Zuordnung:

$L_1 = \{U_1, U_3, U_4\}$	$S_1 = 18$
$L_2 = \{U_5, U_7\}$	$S_2 = 17$
$L_3 = \{U_6, U_8, U_9, U_{10}\}$	$S_3 = 17$
$L_4 = \{U_{11}, U_{12}, U_{15}, U_{13}, U_{16}\}$	$S_4 = 17$
$L_5 = \{U_{17}, U_2\}$	$S_5 = 16$
$L_6 = \{U_{18}, U_{21}, U_{14}, U_{20}\}$	$S_6 = 18$

(U_{19} nicht zugeordnet)

Die Zuordnung ist "falsch gegangen" bei der Wahl von U_2 in L_5 . Wenn dort der Bewerber U_{18} gewählt wäre, würde eine gute Zuordnung gefunden sein. Es ist offenbar wichtig, Elemente mit größerem t_1 (wie U_{18}) eher zu wählen als solche mit niedrigem t_1 (wie U_2). Denn dann bleiben mehr "Scheidemünze" über für

nachher ("öffnen mehrerer Türen in die Zukunft"). Wenn man das Problem 17/18 nochmals bearbeitet mit geänderten P.W.:

$$PW'_i = PW_i + \alpha t_i$$

und dabei versuchsweise $\alpha = 10$ wählt, wird eine Lösung gefunden:

$L_1 = \{U_1, U_3, U_4\}$	$S_1 = 18$
$L_2 = \{U_5, U_7\}$	$S_2 = 17$
$L_3 = \{U_6, U_8, U_9, U_{10}\}$	$S_3 = 17$
$L_4 = \{U_{13}, U_{11}, U_{12}, U_{15}, U_{16}\}$	$S_4 = 17$
$L_5 = \{U_{17}, U_{18}\}$	$S_5 = 18$
$L_6 = \{U_2, U_{21}, U_{14}, U_{20}, U_{19}\}$	$S_6 = 18$

Die ersten vier Arbeitsplätze sind identisch mit denen aus dem vorigen Versuch. Es ist aber zu bemerken, daß schon in L_4 die Folge der Zuordnung sich ändert wegen der anderen Gestaltung des P.W.'s.

Auch für den Fall mit Zonierung und $B_u = 20$ versagt die normale P.W. Methode. Man sieht leicht ein, daß die frühere Wahl einfach zonierter Elemente mehr doppelt zonierte Elemente hinterläßt, also mehr Chancen für die Zukunft läßt. Es ist deshalb angebracht, den P.W. einfach zonierter Elemente zu vergrößern. Zu diesem Zweck sind für die drei Zonen die Mittelwerte m_A , m_B und m_C bestimmt worden der P.W.'s aller Elemente, die die Zonenmarke A, B oder C tragen. Es wurde gefunden: $m_A = 64$, $m_B = 45$, $m_C = 14$. Diese Mittelwerte wurden addiert zu den P.W.'s aller einfach mit A, B oder C zonierte Elemente. Sodann hatte das normale Verfahren, auch mit $B_u = 19$, Erfolg. Das Ergebnis war die Zuordnung der Figur 3.

In den letzten zwei Fällen berücksichtigt der P.W. zwei

heuristische Überlegungen: 1) die Größe der Nachfolgerschar, und 2) der Scheidemünzwert; im letzteren Fall: 1) die Nachfolgerschar, und 2) die Zonierungsindizierung. Es handelt sich hier tatsächlich um eine Abwegung zweier Faktoren. Denn wenn man die zugefügte P.W. Komponente verstärkt (z.B.: αt_i mit $\alpha \gg 10$ wählt, oder das m_A , m_B oder m_C nicht einmal sondern zehnfach zu den PW's der einfach zonierten Elemente addiert) versagt das Verfahren wieder. Es muß betont werden, daß heuristische Argumente nur Plausibilitätswert haben, doch nie zwingend sind. Deshalb wird man immer Gegenbeispiele bedenken können.

7. Monte-Carlo Methoden

Das dynamische P.W. Zuordnungsverfahren kann beschrieben werden mit einem "decision tree", wo jedesmal wenn eine Entscheidung zu treffen ist, diejenige Verzweigung gewählt wird, die den größten P.W. aufweist. Diese Methode ist zwar elegant und einfach, doch saugt sie die Entscheidung in hohem Maße aus den Fingern. Es liegt deshalb nahe, das Zuordnungsverfahren falls Scheitern zu wiederholen (nötigenfalls mehrmalig), wobei die Entscheidungen nicht alle dieselben sein sollen wie vorher. Man kann hierbei systematisch vorgehen und zuerst alle anderen Möglichkeiten bei der letzten Entscheidung versuchen, dann Änderungen bei der vorletzten Entscheidung probieren usw. Da der decision tree meistens sehr ausgedehnt ist, wird diese Methode versagen alleine schon wegen der ungeheuren Administration.

Ein anderer Weg ist, bei den wiederholten Versuchen an jeder Entscheidungsstelle nicht deterministisch zu verfahren, sondern in probabilistischer Weise eine Möglichkeit auszuwählen. Es gäbe an einer solchen Stelle n Möglichkeiten mit den P.W.'s: $PW_1 = PW_2 = \dots = PW_n$. Dann wird die Möglichkeit i ausgewählt mit der Wahrscheinlichkeit p_i , wobei wegen unseres Glaubens am P.W. gelten soll:

$$p_1 \geq p_2 \geq \dots \geq p_n \quad (7.1)$$

Geeignete Formen für die p_i sind:

$$p_i = (PW_i)^\beta / \sum_{j=1}^n (PW_j)^\beta \quad (7.2)$$

oder:

$$p_i = \frac{i}{\gamma} / \sum_{j=1}^n \frac{j}{\gamma} \quad (\text{mit } \gamma = 1) \quad (7.3)$$

Indem man β von ∞ bis 0 oder γ von 0 bis 1 variieren läßt, erhält man eine Skala von Möglichkeiten, sich ausdehnend vom reinen Determinismus bis zur völlig willkürlichen, vom P.W. unabhängigen, Auswahl.

Die P.W. Methode (ohne Zonierung, Schranken 17/18) sei fortgeschritten bis zur partiellen Zuordnung:

$$\begin{array}{ll} L_1 = \{U_1, U_3, U_4\} & S_1 = 18 \\ L_2 = \{U_5, U_7\} & S_2 = 17 \\ L_3 = \{U_6, U_8, U_9, U_{10}\} & S_3 = 17 \\ L_4 = \{U_{11}, U_{12}, U_{15}, U_{13}, U_{16}\} & S_4 = 17 \\ L_5 = \{U_{17}, \dots\} & (S_5 = 13 + \dots) \end{array}$$

Die Bewerber für die fortgesetzte Zuordnung sind:

- a) U_2 $PW_2 = 10$
- b) U_{18} $PW_{18} = 7$
- c) U_{14} $PW_{14} = 5$
- d) U_{20} $PW_{20} = 3$

Wenn jetzt (7.3) mit $\gamma = 0,1$ benutzt wird, erhält man:

$$p_a \approx 0,9 \quad p_b \approx 0,09 \quad p_c \approx 0,009 \quad p_d \approx 0,001$$

Mit 90% Wahrscheinlichkeit wird die falsche Entscheidung a) U_2

getroffen. Bei 10 bis 20 Wiederholungen wird die richtige Entscheidung b) U_{18} sehr wahrscheinlich einmal eintreten.

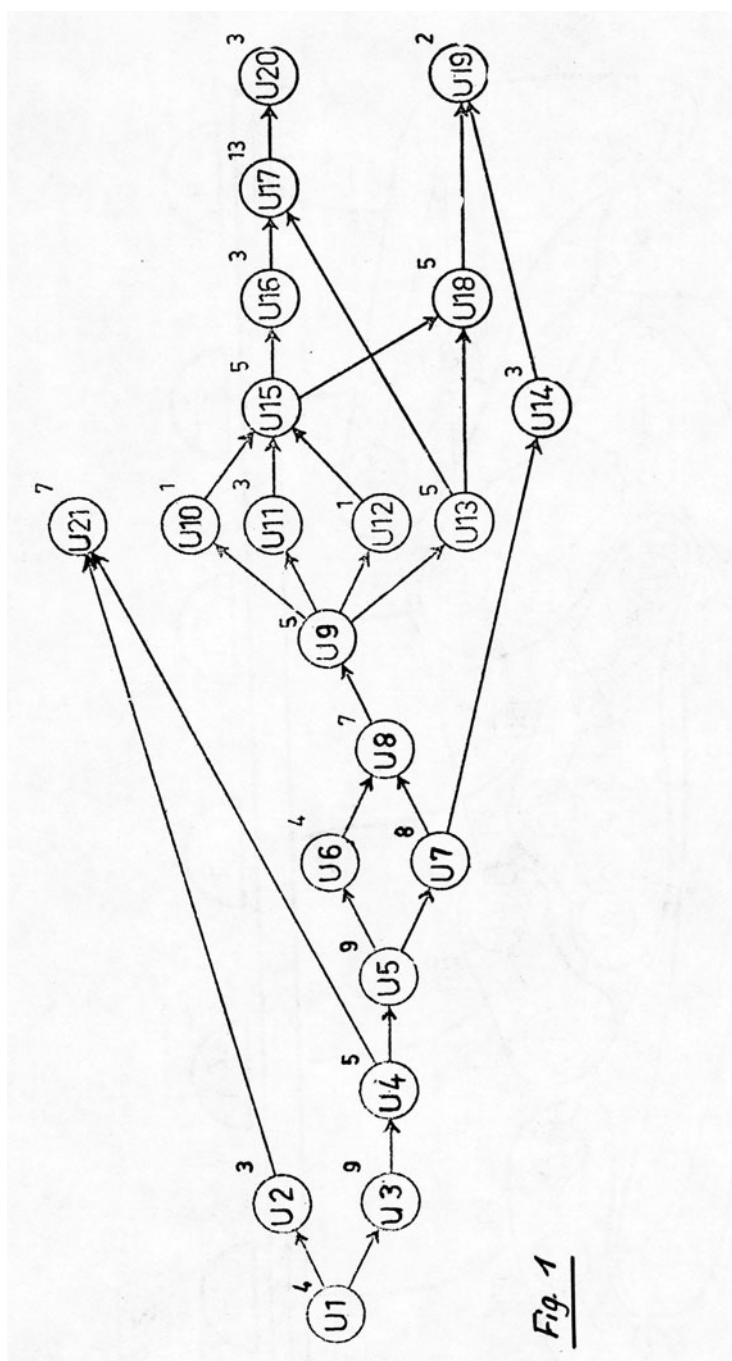
Man kann vielleicht am besten folgendermaßen vorgehen. Zuerst wird $\beta = \infty$ (oder $\gamma = 0$) genommen ($p_1 = 1$, Determinismus). Mißlingt das, dann wird z.B. $\beta = 5$ (oder $\gamma = 0, 1$) genommen, womit einige Zehner der Versuche gemacht werden. Abermals wird bei Scheitern β verkleinert (oder γ vergrößert) und eine neue Versuchsreihe gemacht, usw.

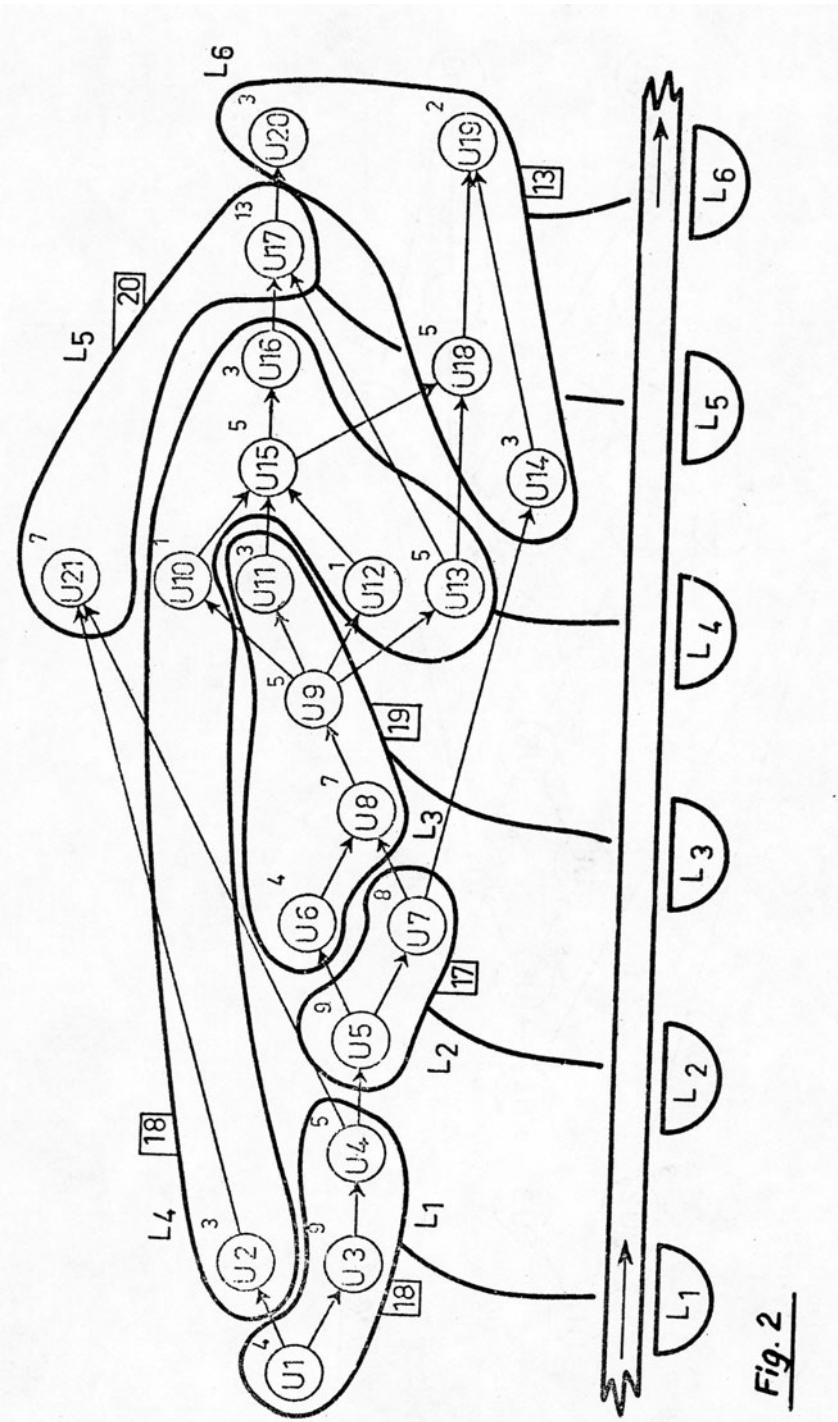
Die Struktur der Figur 1 ist so einfach, daß man die Wahrscheinlichkeit des Erfolges analytisch berechnen kann. Wir haben hier Einfachheit halber an jeder Entscheidungsstelle nur die zwei Bewerber mit größtem P.W. betrachtet und ihnen die Wahrscheinlichkeiten p bzw. $1 - p$ zugeschätzt ($p \approx 1 - \gamma$). In Figur 7 gibt $F(p)$ die Chance $f(p)$ auf Erfolg für den Fall ohne Zonierung mit Schranken 17/18 bei variiertem p . Die Kurve zeigt ein flaches Maximum bei $p = 0,8$ bis $0,85$ (d.h.: $\gamma = 0,15$ bis $0,20$) mit $f(p) \approx 0,22$. Monte-Carlo Versuche ergaben tatsächlich etwa 22% Erfolge. Es braucht kaum betont zu werden, daß die Daten gar nicht normativ sind für kompliziertere Fälle. Es ist bloß zu erwarten, daß das p bei welchem das Maximum auftritt, je größer ist, je besser die Heuristik der P.W.-Evaluation, somit auch das Maximum größer.

Möglichkeiten für Lernverfahren gibt es nicht viel. Wenn das Zuordnungsverfahren scheitert, geschieht das nach vielen Entscheidungen. Das Endergebnis gibt wenig oder gar keine Information über die Ursache des Versagens. Vielleicht die einzige Möglichkeit für Lernverfahren liegt in der automatischen Anpassung des Grades, in welchem die P.W.'s in Abschnitt 6 abgeändert werden.

Anscheinend ist die P.W. Methode versatiler als die Tongesche Methode. Der Grund könnte etwa folgende sein. Das Tongesche Verfahren stützt sich auf Verschiebungen, welche in einer Hierarchie von Elementen vorgenommen werden. Nun impliziert eine derartige Hierarchie eine Affinität benachbarter Ele-

mente, welche in Realität nicht besteht. Dies wird klar, wenn man die in Abschnitt 6 gegebene Optimallösung (mit Schranken 17/18) betrachtet. Arbeitsplatz L_6 enthält U_2 so wie die ganz und gar nicht "benachbarten" Elemente U_{20} und U_{19} . D.h. daß beim Tongeschen Verfahren viele Verschiebungen stattfinden müssen.





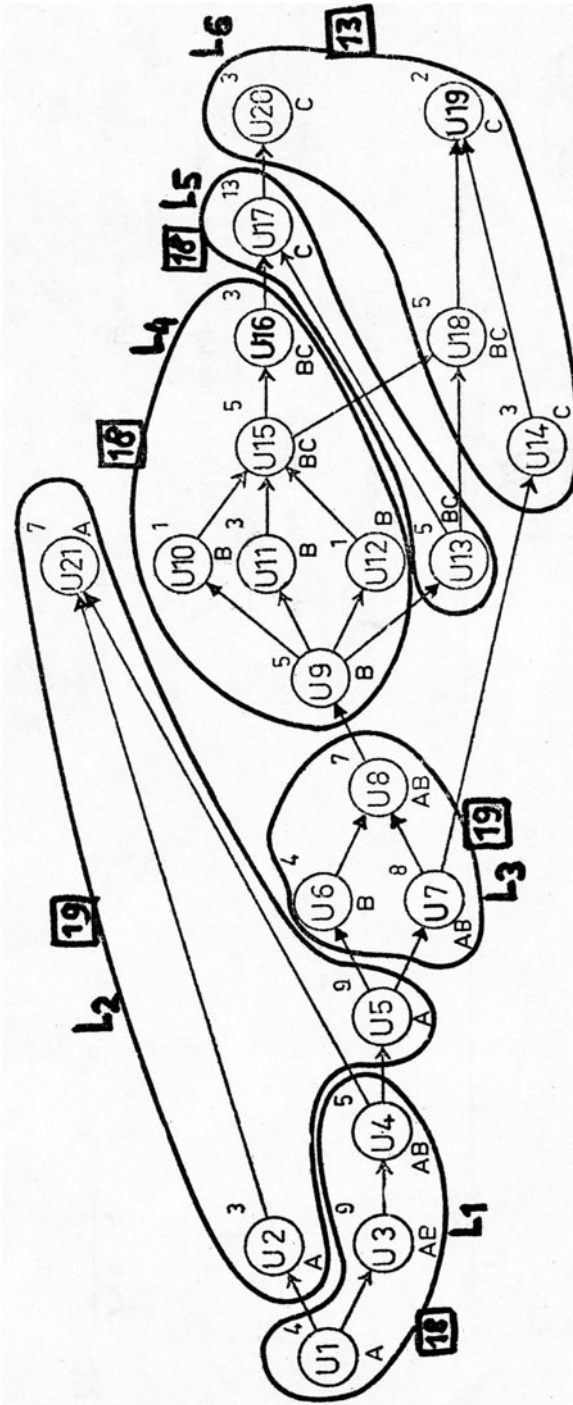


Fig. 3

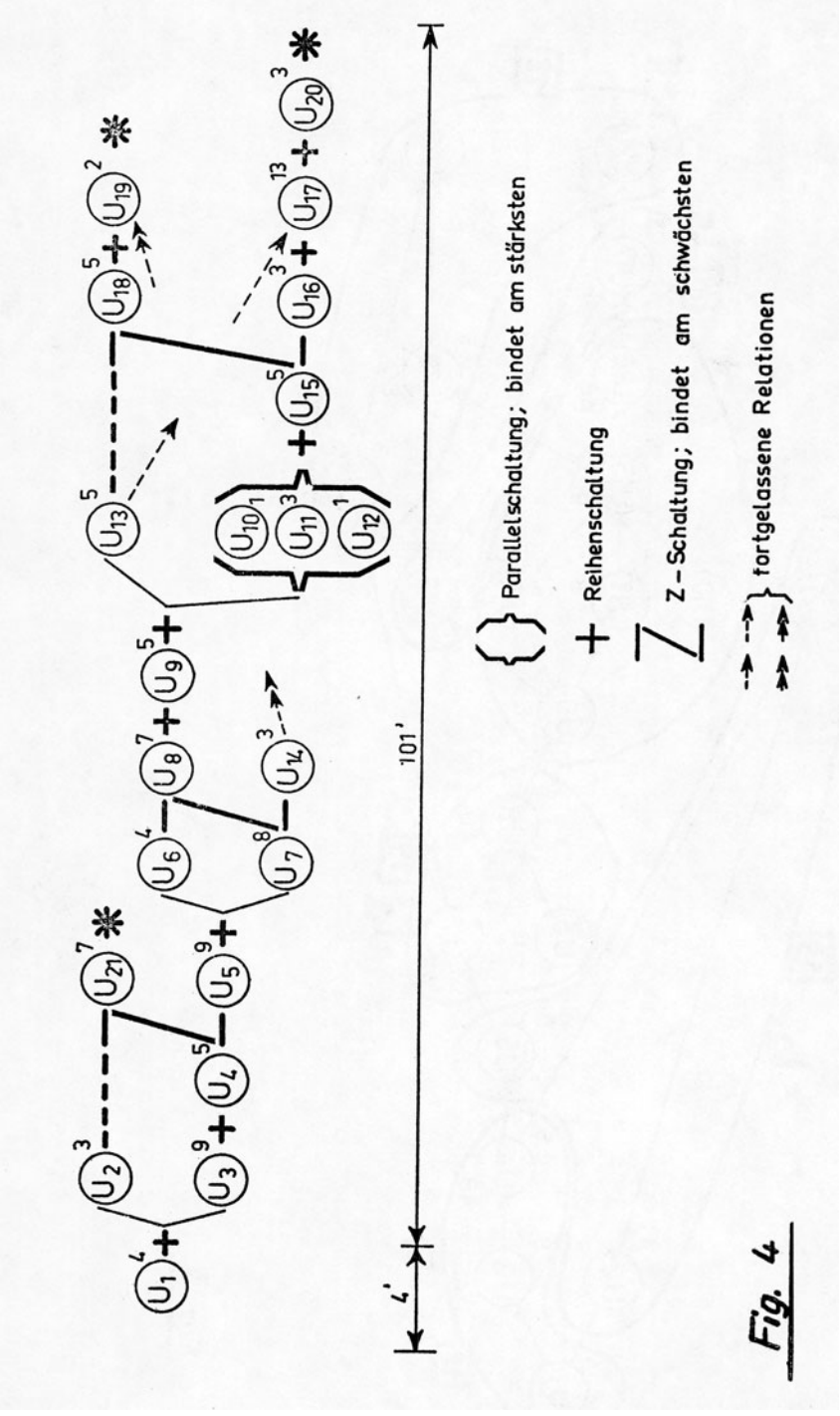


Fig. 4

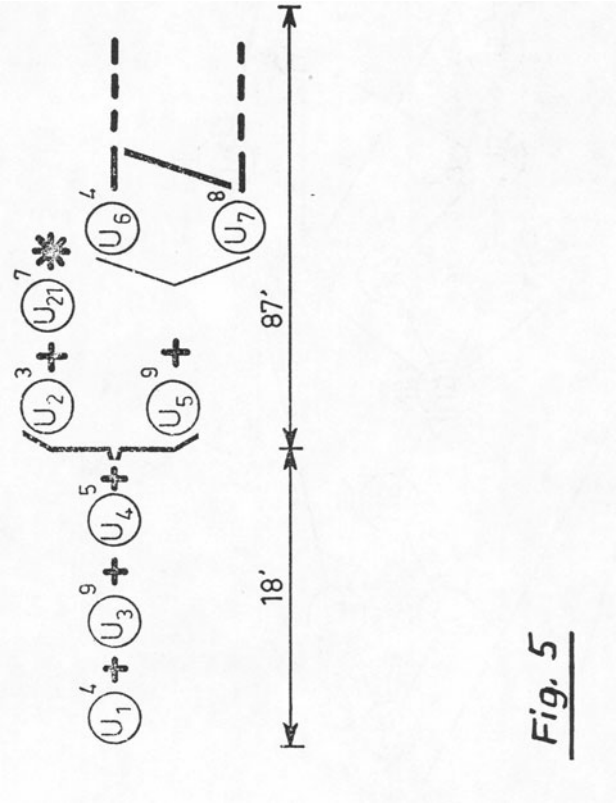


Fig. 5

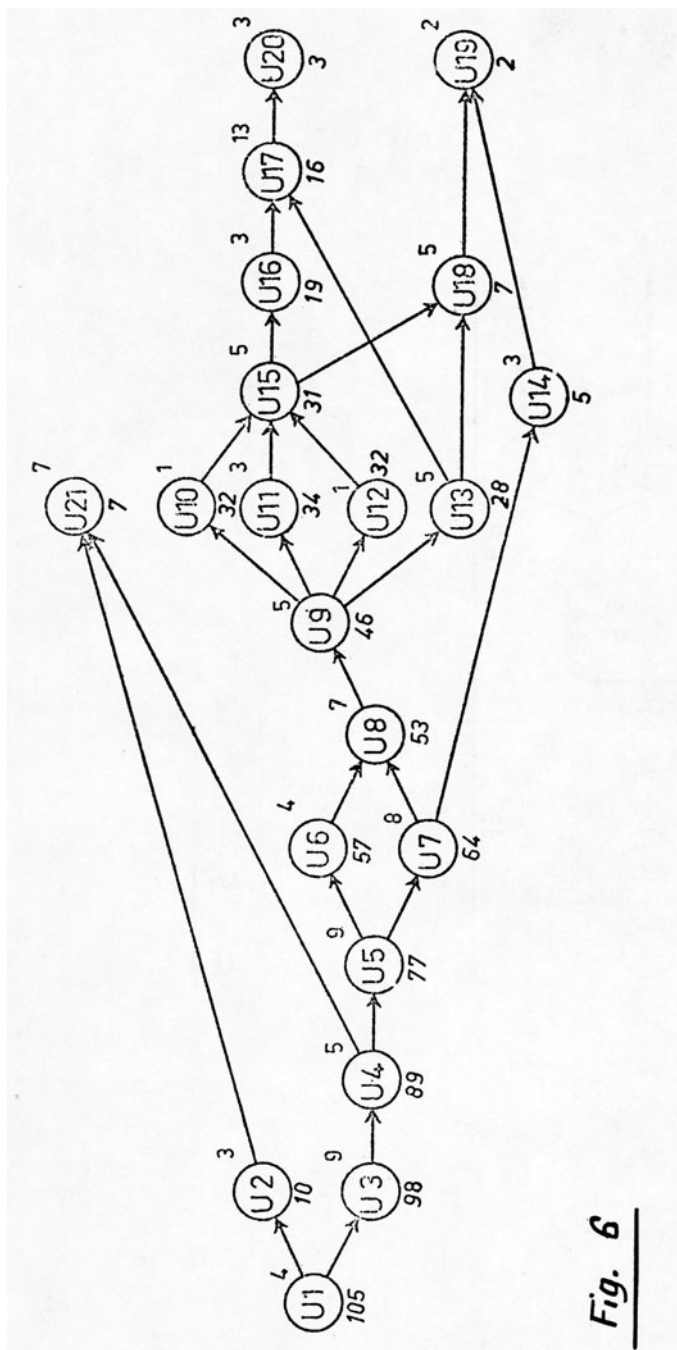


Fig. 6

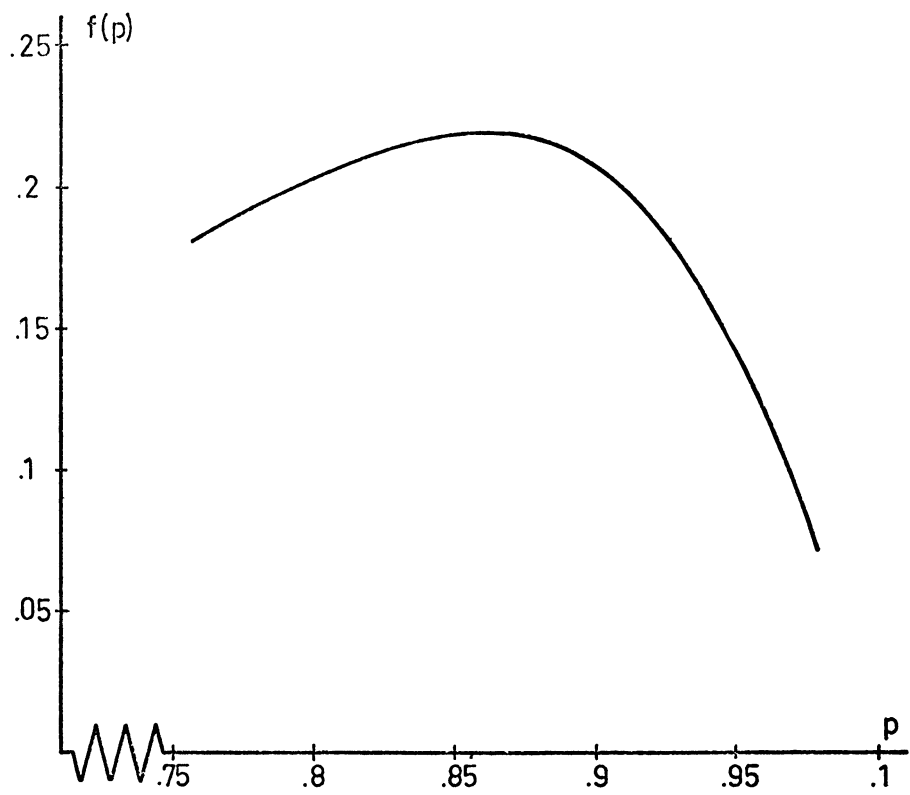


Fig. 7

An Algorithm for the Line Balancing Problem

by R. W. Sierenberg, Delft

In many cases a heuristic method will be the most simple way for solving the assembly-line balancing problem.

However, when the structure of the problem is more complicated (by zoning restrictions or by an overplus of precedence relations) heuristic methods will fail.

Just in these cases the algorithm given here will be helpful.

Logical operations on bitstrings are used in a branch-and-bound method.

In this way it is possible to form and combine a lot of work-stations in a few seconds of computing time.

1. The problem.

The assembly-line balancing problem can be formulated as follows:

given a graph $G = (X, U)$ where X stands for the set of n jobs and U will be the set of precedence relations which specify the permissible orderings of the jobs, so if x_i precedes x_j , $u_{ij} \in U$. G has no circuits, so the jobs can be numbered according to Fulkerson* with the result that if $x_i \rightarrow x_j$ (x_i precedes x_j) then $i < j$.

Figure 1 gives an example of such a graph. To every job $x_i \in X$ a positive real t_i is assigned denoting the execution time of that job. The problem is: find a collection of subsets (work-stations) $(A_1, A_2 \dots A_m)$ of X satisfying the following conditions:

*The numbering procedure of Fulkerson:

- I) In every graph without circuits there will be at least one node without predecessors. Give such a node the next number.
- II) Delete the numbered node from the graph and proceed with I until all nodes are deleted.

- 1) $A_i \cap A_j = \emptyset \quad (i \neq j)$ No job is performed twice or more.
- 2) $\bigcup_{i=1}^m A_i = X$ All jobs are performed.
- 3) $\sum_{x_i \in A_j} t_i \leq c \quad (j = 1 \dots m)$ The sum of the execution-times of the jobs in a work-station should not exceed the cycle time c .
- 4) $\left. \begin{array}{l} \text{if } x_i \rightarrow x_j \\ x_i \in A_p \\ x_j \in A_q \end{array} \right\} p \leq q$ It must be possible to arrange the workstations along the line.
- 5) m is minimized.

At this moment we shall not deal with zoning restrictions. Zoning restrictions make the problem easier according to the method and will be discussed later on.

2. Workstations.

The method consists of two principal parts:

- I Find all feasible workstations with completion times within a given range.
- II Combine as few as possible of those workstations to solve the problem. If this is not possible increase the range mentioned in I and return to I.

It is theoretically possible that this procedure does not lead to a solution with the smallest number of workstations. So one should figure out (with the help of the obtained solution) whether it is necessary to increase the range once again. However, if such a "better" solution exists it is badly balanced. So in practice one tries to find a better fitting cycle time c rather than going on increasing the range too many times.

We shall deal with I first.

According to condition 3 the completion time $t_{A_j} = \sum_{x_i \in A_j} t_i$ of the workstation A_j should not exceed the cycle time c . So c is an upperbound when forming workstations. As a lower bound l we take for instance $l = c - c/10$ initially and decrease l with $c/10$ in the following phases. So computing I we have always a range $[l, c]$ for t_{A_j} .

A workstation A_p is a subset of X but because of condition 4 not every subset of X can be a workstation. Consider for example the subset $\{x_5, x_7, x_8\}$ in figure 1. This cannot be a workstation A_p because x_6 has to be an element of some workstation say A_q .

According to u_{56} $p < q$ must hold but because of u_{68} it has to be $p > q$.

So we need some criterion to figure out whether a job x_p can or can not be joined to a set A to form a workstation. We assume that A itself obeys condition 4.

Furthermore it will be sufficient to formulate the rule only for x_p with p greater than the highest of the indices of the jobs belonging to A , since the workstations will be formed in a dynamic way following the numbers of the jobs as they had been assigned by the method of Fulkerson.

Let Γ_A be the set of all elements $x_q \in X$ satisfying

$$\begin{cases} u_{pq} \in U \\ x_p \in A \\ x_q \notin A \end{cases}$$

and let P_{x_p} be the set of all elements $x_i \in X$ for which there exists a path from x_i to x_p .

Now, we can formulate the rule:

x_p may join the subset A to form a workstation if and only if $\Gamma_A \cap P_{x_p} = \emptyset$.

For, if the intersection is empty there is no path between any element of A and x_p so condition 4 can never be violated. On the other hand, if $x_i \in \Gamma_A \cap P_{x_p}$ then x_i must belong to some workstation $A_j \neq A$ so that $A_j \rightarrow A$ and $A \rightarrow A_j$ should hold at the same time violated condition 4.

3. Bitstrings.

Before we can compute all the workstations mentioned in I, it is necessary to discuss the way to handle sets and operations upon sets in the computer.

Of course it is possible to use a series of memory cells (array) and fill the i -th cell with a one or a zero according to whether element x_i belongs to the set or not.

Operations on sets will lead to the comparison of arrays cell by cell. If the method is programmed in this way, the execution time of the program will be prohibitive. As a matter of fact the leading idea of the method is to use bitstrings to represent sets. So we use the bits of the memory cells instead of the cells themselves.

The workstation $\{x_2, x_6, x_8, x_{14}\}$ (fig. 1) for instance is represented by 01000101000001000000 being the first 21 bits of a memory cell (the remaining bits are supposed to be 0).

Taking the union or the intersection of two sets results in logical operations upon cells. Every assembly-language contains instructions for this kind of operations.

Mostly they are only apparently represented in a higher level language or not at all. So these languages have to be supplemented by a number of subroutines written in assembly-language.

When this is done, these subroutines will not only be helpful to solve the line balancing problem but be a very important tool for solving other combinatorial problems such as scheduling problems, transportation problems and the construction of time tables.

We shall use 3 logical operations:

1. The intersection operation \cap defined by

$$\begin{array}{r} 1010 \\ 0110 \\ \hline 0010 \end{array} \cap$$

2. The union operation \cup defined by

$$\begin{array}{r} 1010 \\ 0110 \\ \hline 1110 \end{array} \cup$$

3. The logical sum \oplus defined by

$$\begin{array}{r} 1010 \\ 0110 \\ \hline 1100 \end{array} \oplus$$

A set is empty if the numerical value of the representing cell is zero. This can be tested in an easy way.

Obviously when the number of elements of a set is greater than the number of bits in a cell, we have to use more cells to represent it. This will not necessarily increase the run time very much because we shall mostly be concerned with finding a special bitstring in a series of bitstrings or finding one which has an empty intersection with a given one.

If we are in luck we can see after testing only one part of a string that this string is not a good one.

Moreover we can help ourselves to be in luck by testing first that part of the string, where the test failed in the same part of its neighbour; since the bitstrings are never ordered in a random way.

4. Algorithm to find workstations.

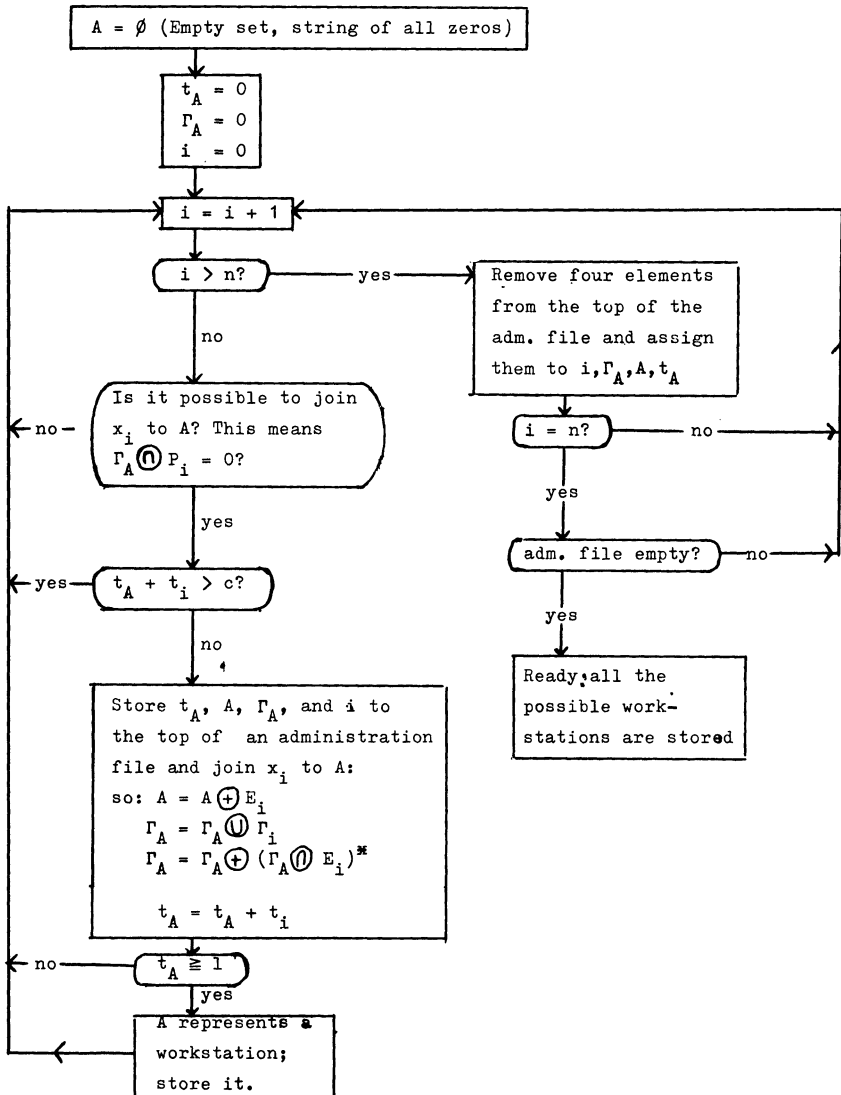
Use is made of the rule formulated in 2. Therefore it is convenient to form three bitstrings for each job x_i :

- 1e. E_i : a string where only bit i is set;
- 2e. P_i : a string representing the set P_{x_i} (see 2);
- 3e. Γ_i : a string representing the set Γ_{x_i} of elements x_q with $u_{iq} \in U$.

Γ_A defined in 2 will be considered as a bitstring too.

Again we repeat that the jobs are assumed to be numbered in the way of Fulkerson.

It assures us that all workstations can be found while the number of tests is as few as possible.



Algorithm to find workstations.

* This statement clears bit i in Γ_A whether it was set or not.

If we use, for example, the algorithm for the problem given in fig. 1 with $c = 18$ and $l = 17$, we get 153 possible workstations; we show here the first and the last ten:

nr.	indices of the jobs				
1	1	3	4		
2	2	3	4		
3	2	4	5		
4	2	6	7	14	
5	2	6	8	14	
6	2	6	14	21	
7	2	7	8		
8	2	7	21		
9	2	8	9	10	12
10	2	8	9	11	
.					
.					
.					
144	14	16	18	21	
145	14	17	19		
146	14	18	19	21	
147	14	18	20	21	
148	15	18	21		
149	15	17	19		
150	16	18	19	21	
151	17	18			
152	17	19	20		
153	18	19	20	21	

Now we have to combine a number of these workstations to solve the entire problem.

But before doing so, we first examine the list because there are workstations for which we can figure out in advance that they will never occur in a solution.

To inspect this we form a bitstring Q_{A_i} for each workstation. Bit j is set if there either exists a path from $x_j \notin A_i$ to some $x_p \in A_i$ or a path from some $x_p \in A_i$ to $x_j \in A_i$.

For every $x_j \in Q_{A_i}$ there must exist a workstation A_p with $A_p \cap A_i = \emptyset$ and $x_j \in A_p$.

If not it is impossible for A_i to occur in a final solution.

Consider for instance $A_2 = \{2, 3, 4\}$

Then $Q_{A_2} = \{1, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21\}$. x_1 only occurs in A_1 but $A_1 \cap A_2 = \{3, 4\} \neq \emptyset$.

So A_2 can be removed from the list.

It turns out that in the example the 153 workstations can be reduced to 90 in this way.

5. Algorithm to combine workstations.

Similar to Gutjahr and Nemhauser [1] we shall define a state S as a set of disjunct workstations but so that if a job x_1 belongs to S all the predecessors of x_1 belong to S .

When we combine workstations we shall follow the strategy that at every moment the combined workstations form a state.

Let P_{A_i} be the bitstring denoting the set of all predecessors of A_i . A_i may join the combination S if $A_i \odot S = 0$ (The elements of the new workstation must be disjunct from S) and if $P_{A_i} \oplus (P_{A_i} \odot S) = 0$ (S must contain all the predecessors of the new workstation).

The general idea of this algorithm is as follows:

1. $S = 0$;
2. Find the next workstation in the list of workstations that can be joined to S ;
3. If there is no such workstation, delete the workstation from S that was joined to S the last time and go on with 2;
4. If there is such a workstation, join it to S and go on with 2 until all solutions are found.

Because of the requirement that S must contain all the predecessors of a workstation to join it to S , it can happen that a workstation cannot be joined to S until another one is joined to S that is further down the list.

This makes it necessary to inspect the list cyclically.

In order to increase the efficiency of the algorithm the workstations are grouped in classes.

Class 1 contains all the workstations containing x_1 ;
 class 2 contains all the workstations containing x_2 but not x_1 ;
 In general class j contains all the workstations containing x_j but not those with x_i | $i < j$.

So the classes are disjunct. Some may be empty. The algorithm described in 4 already yields the workstations in the desired sequence.

We use the classes as follows:

if we enter a class k , searching for the next workstation, we can skip this class if bit k is set in S (because $A_i \cap S$ must be \emptyset).

For the same reason we can skip the rest of a class when we just joined a workstation of that class to S .

In a simple example as the one we use here, it follows by a simple calculation that the minimum number m of workstations that we have to combine to form a solution must be 6 ($l = 17$, $c = 18$). In general however, this will be not as easy. But it is always possible to estimate m . If there exist solutions with a smaller m , m will be decreased by the algorithm. But the algorithm will never deal with solutions containing more than m workstations.

m is used as follows:

If at any time S consists of $m-1$ workstations and S itself is not a solution, we form a bitstring R denoting the jobs that are missing in S for S to be a solution.

There exists a solution with m workstations containing S only if there exists a workstation A_i equal to R .

This A_i must be found in the class denoted by the lowest numbered bit in R .

As we described above, it is necessary to inspect the list cyclically. This causes the need of some mechanism to avoid that a possibility is tried out twice or more.

Say we did inspect all the combinations with A_i and A_j . So we have to delete A_j from S . Let A_k ($k > j$) the next workstation that is joined to S after deleting A_j .

By searching cyclically it could happen that A_j is joined to S again. But we inspected the combination $A_i - A_j - A_k$ before.

To avoid these and similar difficulties we set bit 0 in A_j at the moment it is deleted from S . If we take care that bit zero is always set in S it is impossible to join A_j to S once more. Bit 0 must be reset in A_j at the moment A_i is deleted from S , in order to admit combinations with A_j not already inspected.

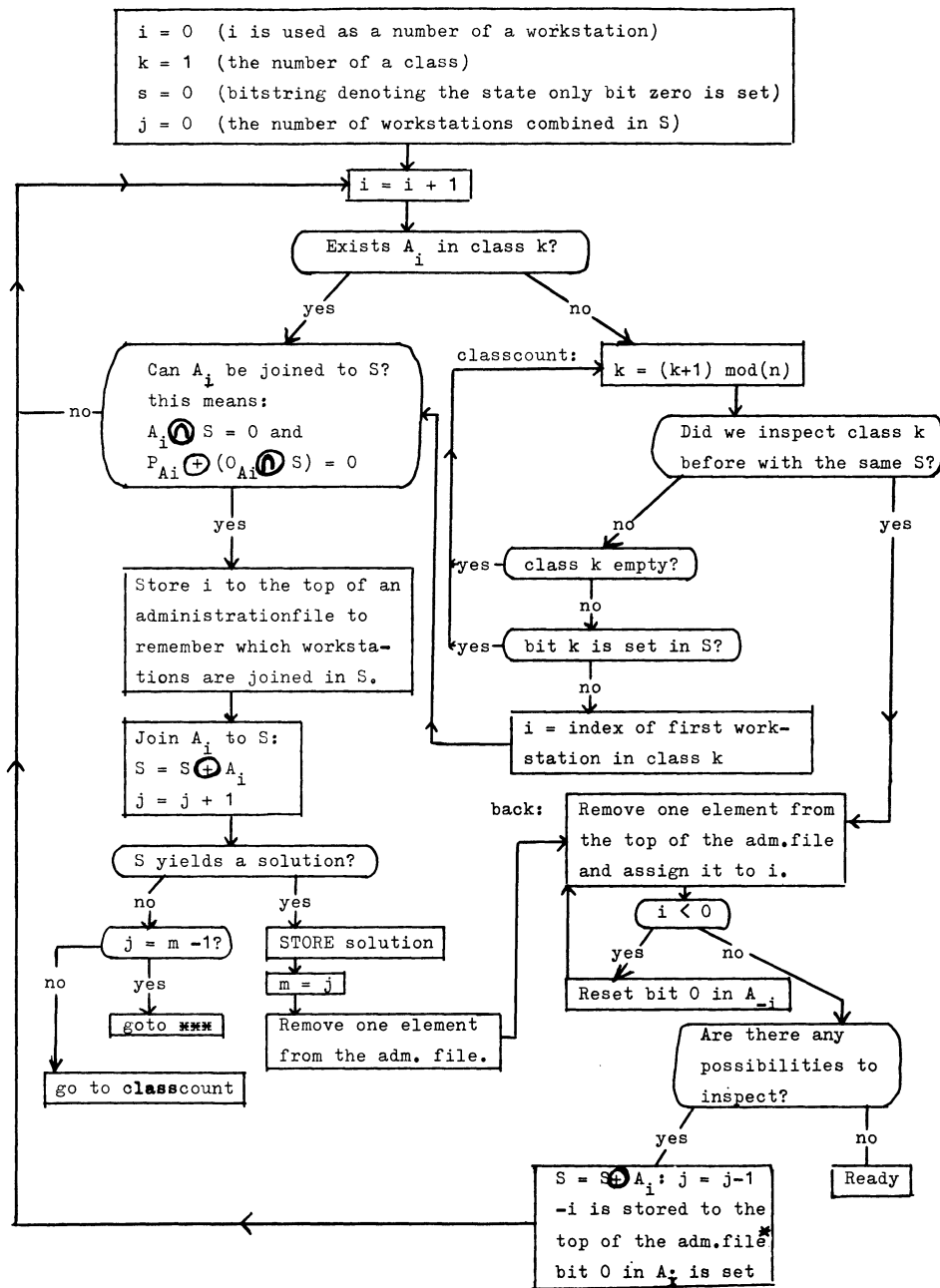
This completes the description of the algorithm. The flowchart is given on the next page.

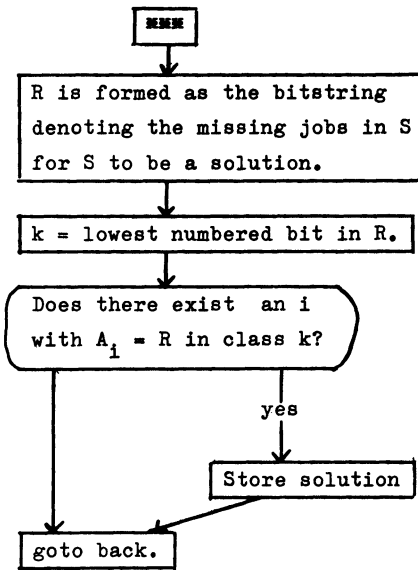
There are 30 solutions of this simple example.

They are found with the help of a medium size computer (TR4) within one second of computertime (deleting in- and outputtime).

We choose this example (due to Tonge [7]) to make a comparison possible with the heuristic methods described by Kosten [5].

If it turns out that there are so many solutions it is evident that a heuristic method will be succesful. When we, however, deal with zoning restrictions the utility of this method will be more evident.





Flowchart of the combination program.

* -i is stored to the top of the administrationfile to remember that bit zero is set in A_i .

Figure 2 shows solution number 15 as an example of a solution with $c = 18$ and $l = 17$.

(If we want to know a best balanced solution of this example we have to figure out if there for instance exist solutions with $l = c = 21$ ($m = 5$). It turns out that there are 8 solutions. figure 3 shows one of them).

6. Zoning restrictions.

Now we shall deal with zoning restrictions.

This means that for every job one or more physical situations are specified for the performance of that job.

All the jobs that can be performed in the same situation of the line form a zone. So there are as many zones as there are different situations of the line.

The zones need not to be disjunct.

We shall say that a zone is singular if the situation the zone represents can occur just one time along the line.

The most simple case is:

every job belongs just to one zone and every zone is singular.

In this case every zone can be treated as a subproblem.

Every combination of solutions of those subproblems solves the entire problem.

Figure 4 shows a more complicated example (due to Tonge [7]).

Jobs can belong to more than one zone and some zones are not singular (zone 2 and 3). Here we see that zone 3 is not singular because for instance x_{13} and x_{33} belong to zone 3, but x_{23} does not belong to zone 3 and is situated between them. The situation denoted by zone 3 must be repeated along the line.

Most of the zones are singular. These zones can be considered as subproblems. Fig. 5 gives the subproblem for zone 1.

We can use the method discussed above to find all possible solutions of such a subproblem.

Here a solution need not contain all the jobs but only those that must belong to the zone.

(This new definition of a solution and the requirement to find all the solutions - so not only those which minimize n - cause some simple modifications in the combination program).

A zone that is not singular cannot be treated as a whole because some precedence relations might be violated. In these cases we add dummy jobs to the zone in such a way that they show how the other zones are situated in that zone.

We give a dummy job a performance time equal to c to avoid that a dummy is joined with the normal jobs in workstations. Furthermore we shall say that a dummy must belong to a solution to avoid a number of solutions with the only difference that a dummy is or is not represented in it.

Now that we have added these dummies (see fig. 6 for an example of zone 3) we can also treat a zone that is not singular as a whole. In the solutions the dummies are deleted again.

If we are finished with all the zones in this way, every zone gives a list of solutions. Now we have to choose one element of each list, in such a way that the chosen ones together solve the entire problem. The algorithm to perform this is already available. It is the combination algorithm again where we take the lists of solutions instead of the classes of workstations.

Figure 7 shows an entire solution with $c = 156$. This is not quite legal because of u_{15} , u_{16} and u_{18} for which we changed the execution times to 50, 100 and 156. Otherwise the cycletime c should be 319 which makes the problem almost trivial.

References.

1. A.L. Gutjahr & G.L. Nemhauser An algorithm for the line
balancing problem,
Management Science vol. 11, no. 2
Nov. 1964.

2. M. Held, R.M. Karp & R. Sharestian Assembly line balancing-dynamic
programming with precedence constraints
Operations Research vol. 11, 1963.

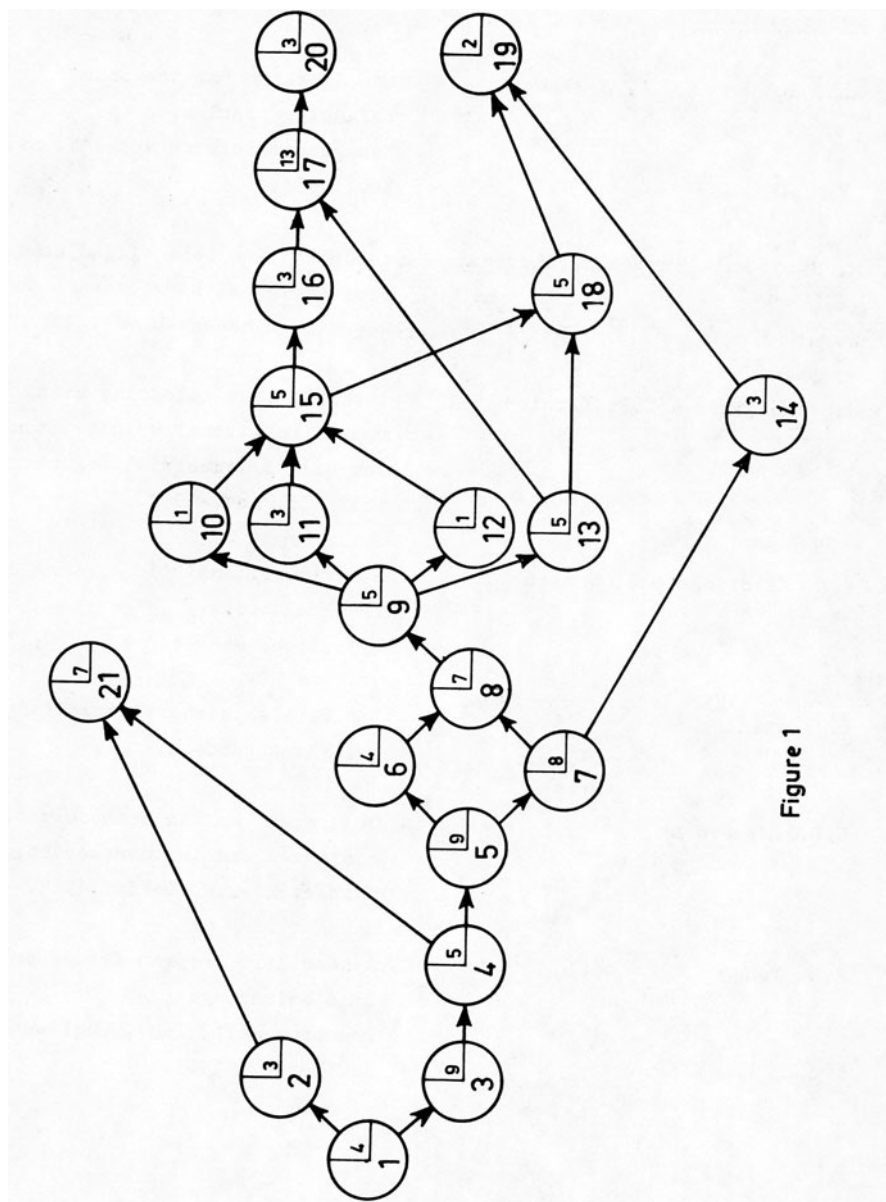
3. W.B. Helgeson & D.P. Birnie Assembly line balancing using the
ranked positional weight technique.
Journal of Industrial Engineering,
vol. XII, nov. 1961.

4. N.D. Kilbridge & L. Wester A review of analytic systems of
line balancing.
Operations Research vol. 10, 1962.

5. L. Kosten Das Problem der Arbeitszeitangleichung
bei Fließbändern.

6. L.B.J.M. Sturm Onderzoek van een methode voor het
afstemmen van lopende banden,
Master's th. T.H. Delft, 1968.

7. F.M. Tonge A heuristic program for assembly
line balancing.
Prentice Hall, Inc., Englewood Cliffs,
New Jersey, 1961.



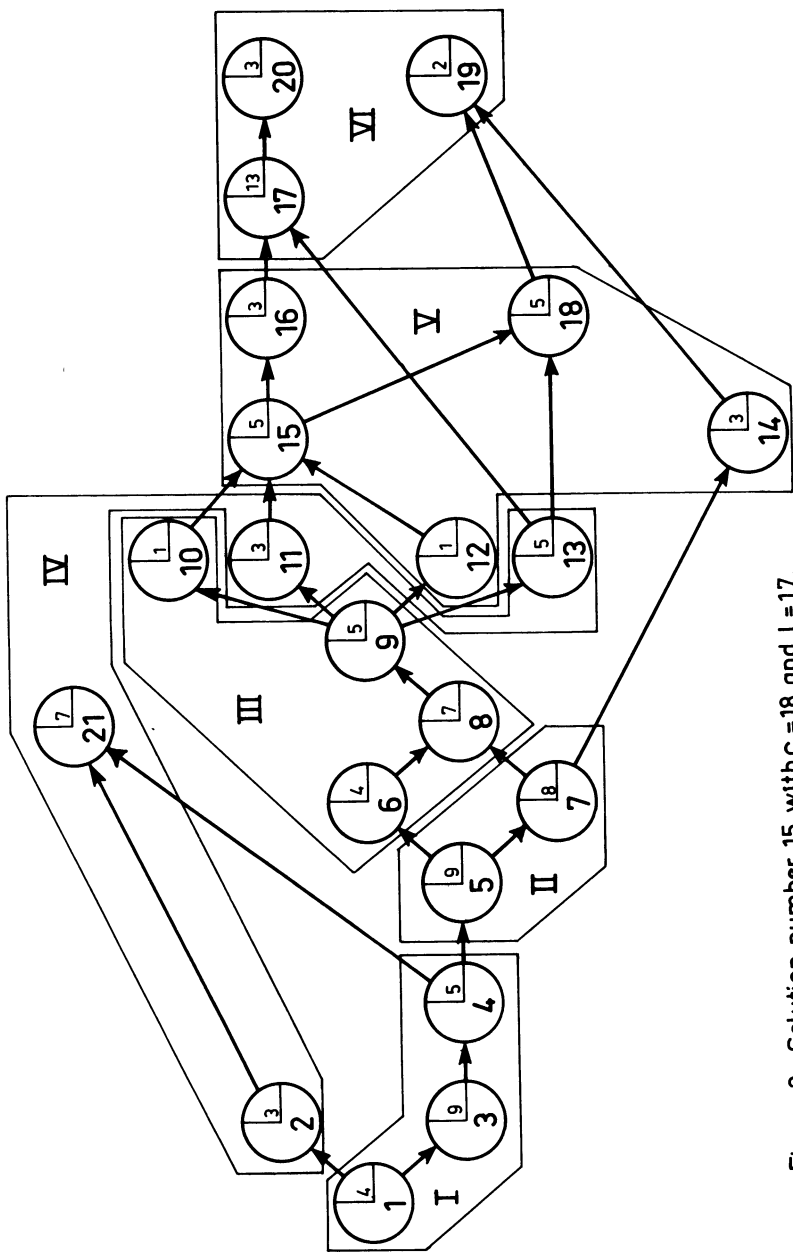


Figure 2. Solution number 15 with $C = 18$ and $l = 17$.

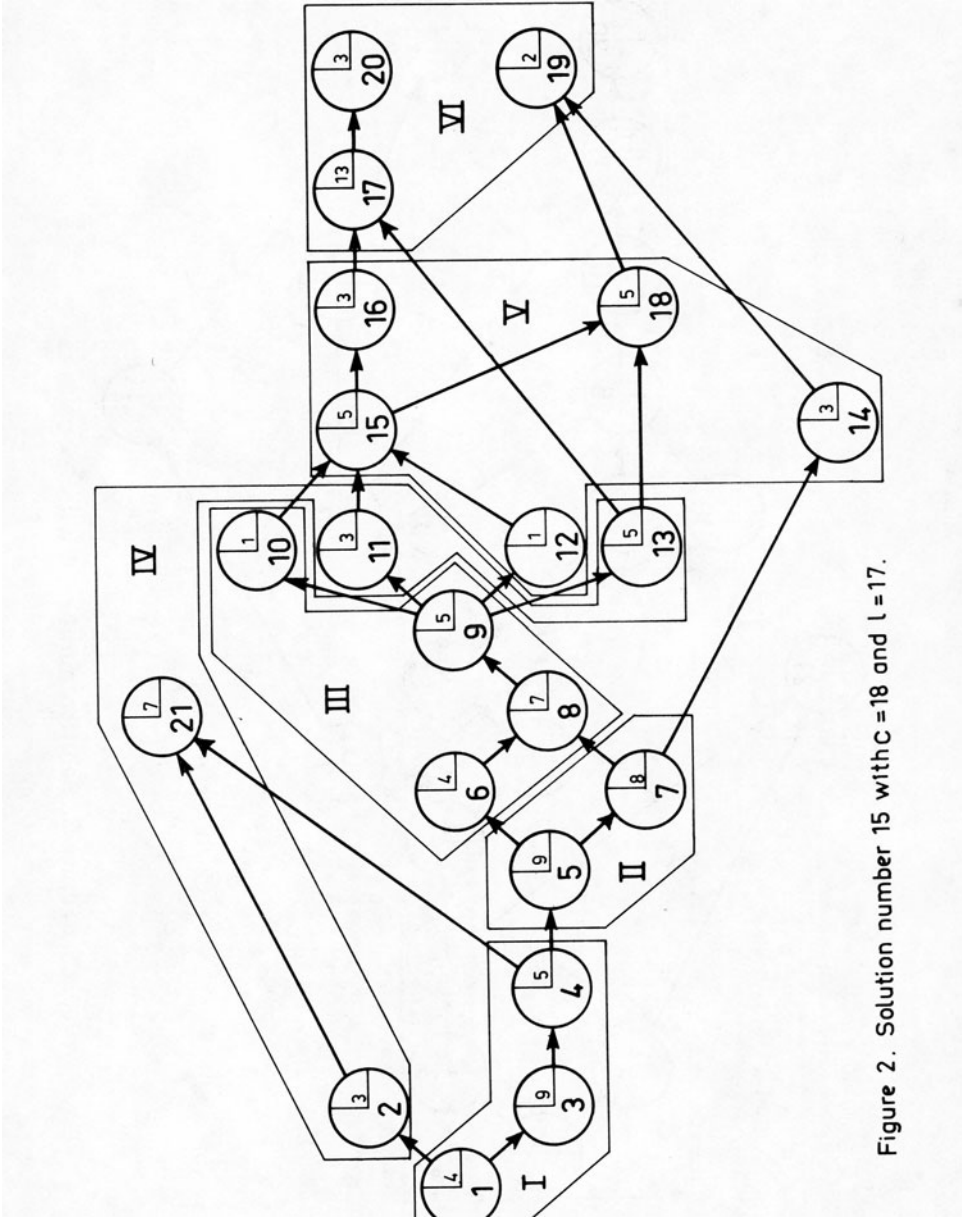
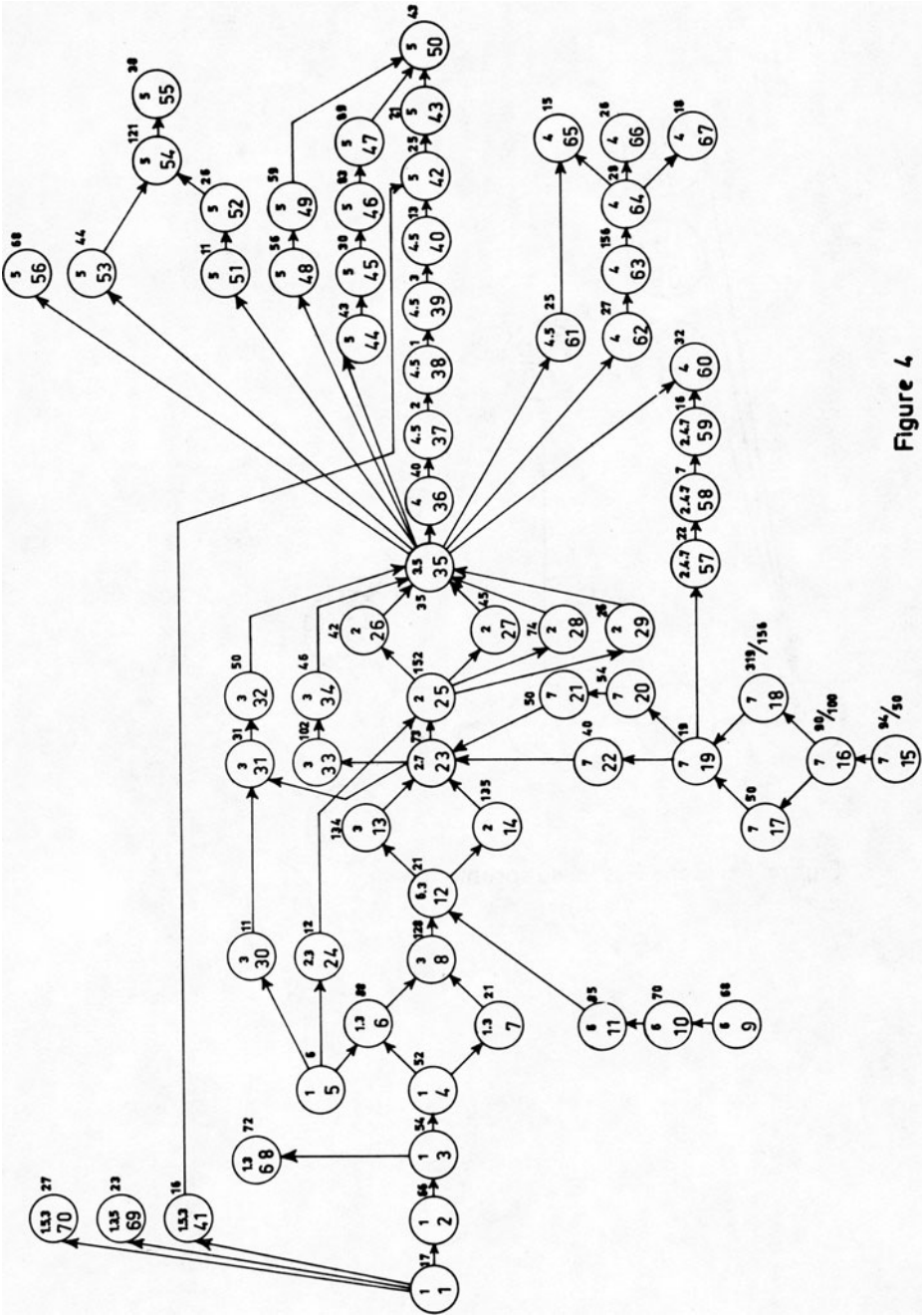
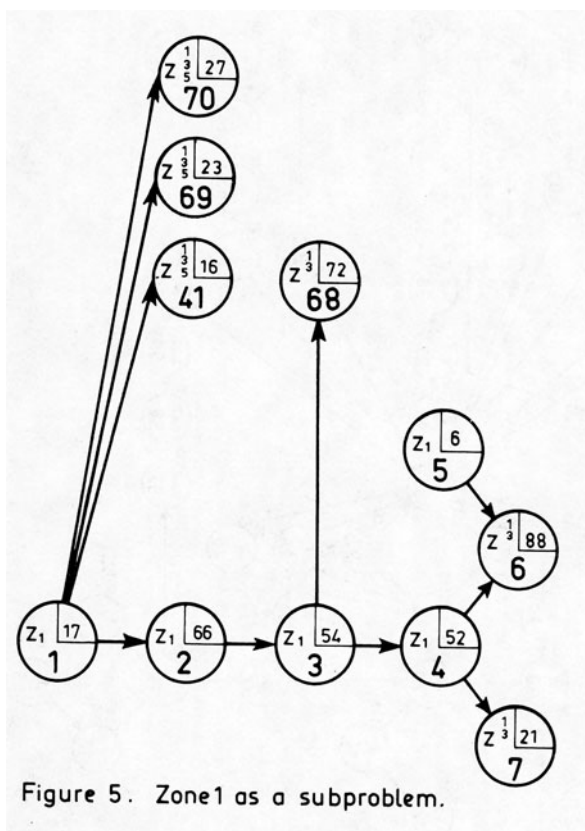


Figure 2. Solution number 15 with $c=18$ and $l=17$.





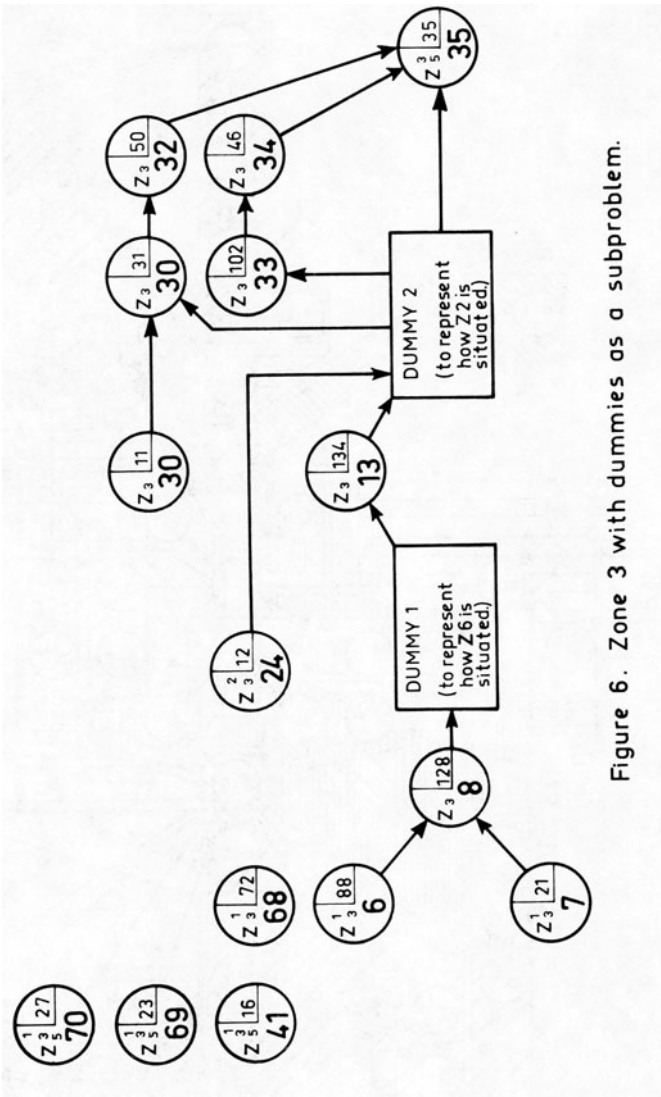


Figure 6. Zone 3 with dummies as a subproblem.

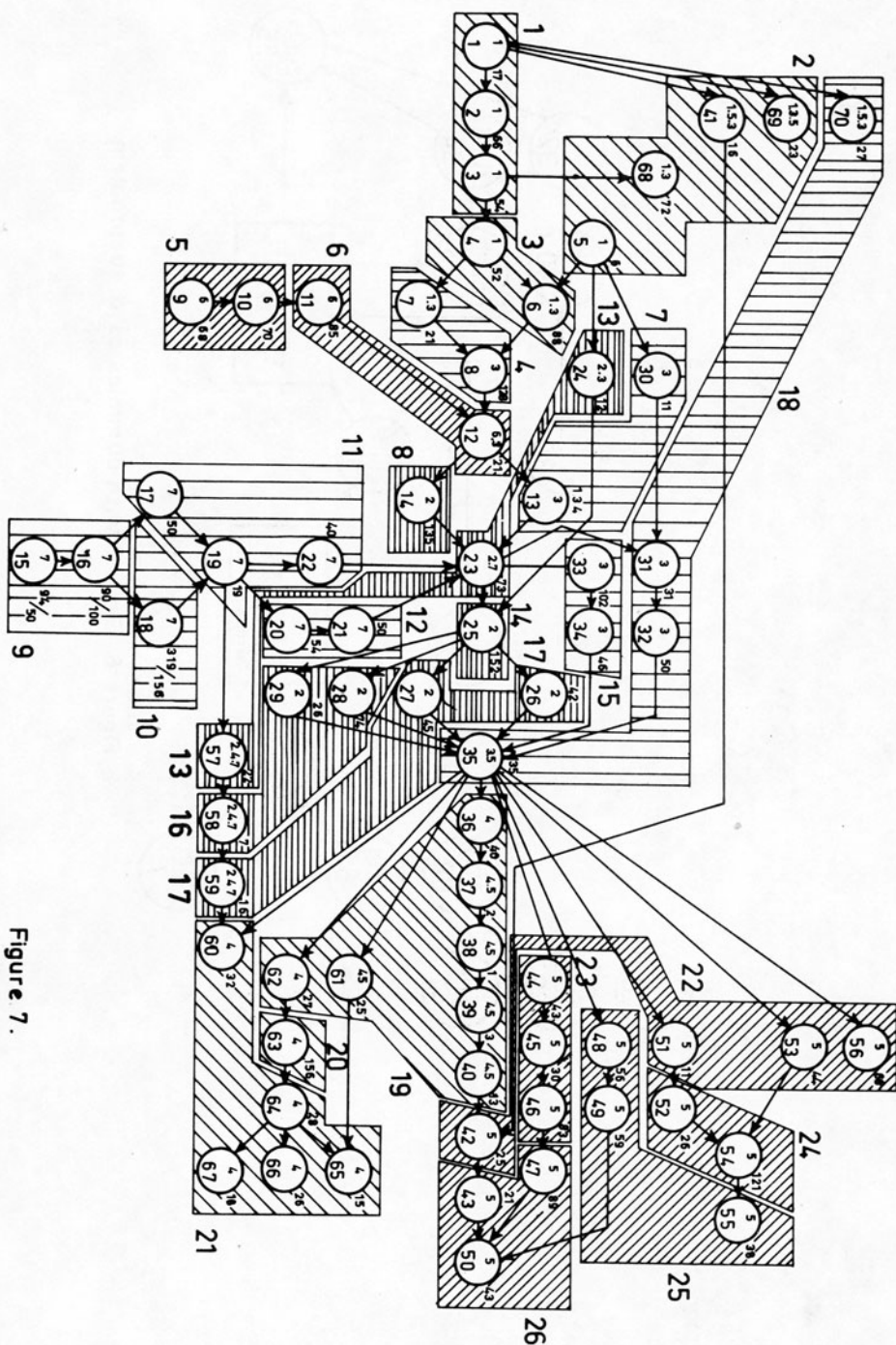


Figure 7.

Ganzzahlige Programmierung

Modifikationen von Cutting-Plane-Methoden der ganzzahligen Optimierung

von H. Müller-Merbach, Darmstadt

0. Kurzfassung oooooooooooooooo

In diesem Beitrag werden ein Überblick über die aus der Literatur bekannten Cutting-Plane-Methoden der ganzzahligen Optimierung gegeben und einige Vorschläge zur praktischen Verbesserung dieser Methoden gemacht.

Die Methoden zur Erzeugung von Cutting-Planes werden in die "direkten" und "indirekten" Methoden gegliedert. Als "direkte" Methoden werden solche bezeichnet, die aus jeder beliebigen Nebenbedingung durch "ganzzahlige Division" Cutting-Planes erzeugen. Dagegen werden als "indirekt" solche Methoden bezeichnet, die über den Umweg des nichtganzzahligen Optimums Cutting-Planes erzeugen.

Die in diesem Beitrag vorgebrachten Vorschläge betreffen folgende Punkte:

- Vor Beginn der Berechnung des ganzzahligen Optimums kann es oft nützlich sein, mit Hilfe von heuristischen Verfahren gute Näherungslösungen zu bestimmen.
- Ferner lassen sich ebenfalls vor Beginn der Berechnung des ganzzahligen Optimums viele direkte Cutting-Planes erzeugen, deren restriktivste dem Problem von Beginn an hinzugefügt werden sollten.

- Wenn man dann mit der Simplex-Methode das nichtganzzahlige Optimum berechnet und von dort mit indirekten Cutting-Plane-Methoden weiterarbeitet, dann sollte man pro-Iteration nicht jeweils nur einen Cutting-Plane hinzufügen, sondern möglichst mehrere.
- Diese indirekten Cutting-Planes sollte man aber durch die direkten Cutting-Plane-Methoden noch verschärfen, was in den meisten Fällen möglich ist.

1. Allgemeines

1.1. Stand der Entwicklung in der ganzzahligen Optimierung

Die Lösung von Optimierungsproblemen mit Nebenbedingungen bereitet nach wie vor unüberwindliche Schwierigkeiten, wenn für einige oder alle Variablen ganzzahlige Werte gefordert sind und die Zahl dieser "ganzzahligen" Variablen größer als - je nach Problemstruktur - 30 bis 100 ist.

Die ältesten und elegantesten Methoden zur Lösung derartiger Probleme sind die Cutting-Plane-Methoden, wie sie von GOMORY (1958), DANTZIG (1959), GOMORY (1960), GOMORY (1963) und neuerdings von YOUNG (1968), GLOVER (1968), BALAS (1971), BALAS, BOWMAN, GLOVER, SOMMER (1971) sowie MÜLLER-MERBACH (1971b) entwickelt wurden. Bei ihnen wird der durch die Restriktionen gegebene zulässige Lösungsbereich iterativ durch Einfügen von "Cutting-Planes" (Schnittebenen) verkleinert, bis das z. B. mit der Simplex-Methode gefundene Optimum alle Ganzzahligkeitsbedingungen erfüllt. Die Cutting-Planes werden so gebildet, daß keine zulässigen ganzzahligen Punkte, sondern nur nicht-ganzzahlige Bereiche ausgeschlossen werden.

Neben den Cutting-Plane-Methoden gibt es enumerative Methoden (Branching and Bounding, begrenzte Enumeration etc.), die auf LAND und DOIG (1960), DAKIN (1963), BALAS (1965) u. a. zurückgehen. Hier wird der zulässige Lösungsbereich vom nicht-ganzzahligen Optimum aus systematisch nach ganzzahligen Lösungen abgesucht.

Weder die Cutting-Plane-Methoden noch die enumerativen Methoden garantieren bei großen Problemen das Finden der optimalen Lösung in einer angemessenen Rechenzeit. Aus diesem Grund wurden - besonders in den letzten Jahren - Näherungsmethoden entwickelt, mit denen man mit relativ niedrigem Rechenaufwand oft eine gute, wenn

auch nicht immer die optimale Lösung findet. Solche Methoden sind u. a. von ECHOLS und COOPER (1968), HILLIER (1969), KREUZBERGER (1969a, 1969b, 1970) und MÜLLER-MERBACH (1969, 1970, 1971a, S. 391-395) beschrieben worden.

Da die Näherungsmethoden nicht mit Sicherheit zum Optimum führen, wird in vielen Forschungsgruppen an der Verbesserung der exakten Methoden gearbeitet. Die Fortschritte sind zwar gering, nichtsdestoweniger jedoch spürbar.

In diesem Beitrag sollen einige Modifikationen der Cutting-Plane-Methoden vorgetragen werden, die bei Problemen, in denen alle Variablen der Ganzzahligkeitsbedingung unterliegen, gewisse Beschleunigungen im Lösungsablauf bewirken können.

Ein wesentlicher Teil der hier behandelten Modifikationen besteht in der Kombination von "direkten" und "indirekten" Cutting-Planes. In den Abschnitten 2 und 3 werden diese beiden Typen von Cutting-Planes und ihre bekanntesten Formen repetiert. Dabei seien solche Cutting-Planes als "direkt" bezeichnet, die aus einer Beziehung $x_{n+1} + \sum_{j \in N} a_{ij} x_j = a_{i0}$ durch "ganzzahlige Division" der Koeffizienten a_{ij} hergeleitet werden. Dagegen seien als "indirekt" solche Cutting-Planes bezeichnet, die sich durch Trennen der ganzzahligen und gebrochenen Teile der Koeffizienten ergeben, einschließlich der Intersection Cuts. Direkte Cutting-Planes lassen sich "direkt" aus jeder Restriktion einer beliebigen Lösung herleiten. Indirekte Cutting-Planes werden im allgemeinen von den Restriktionen der Optimallösung hergeleitet; aus Restriktionen der Ausgangslösung sind gewöhnlich keine indirekten Cutting-Planes zu erzeugen. Zu den direkten Cutting-Planes gehören alle solchen, die auch unter Namen wie "all-integer" bekannt sind, also die von GOMORY (1963), YOUNG (1968), GLOVER (1968). Dagegen zählen zu den indirekten Cutting-

Planes die von GOMORY (1958), DANTZIG (1959), GOMORY (1960), BALAS (1971), BALAS, BOWMAN, GLOVER, SOMMER (1971) und MÜLLER-MERBACH (1971b).

Im Abschnitt 4 werden dann die Möglichkeiten einer Verbesserung der Cutting-Planes durch die Kombination von "direkter" und "indirekter" Methode erörtert. Ferner werden Kriterien angegeben, nach denen unter den vorhandenen Cutting-Planes die k "besten" ausgewählt werden können.

1.2. Ein Zahlenbeispiel xxxxxxxxxxxxxxxxxxxxxxxxxxxx

Zur Demonstration der einzelnen Methoden wird das folgende Beispiel verwendet:

Maximiere z

$$\begin{aligned} z - 7x_1 + x_2 &= 0 \\ 16x_1 - 12x_2 &\leq 41 \\ 4x_1 + 4x_2 &\leq 43 \\ -4x_1 + 4x_2 &\leq -5 \end{aligned}$$

$$x_1, x_2 \geq 0, \text{ ganzzahlig}$$

Nach Hinzufügen der Schlupfvariablen x_3 bis x_5 , die ebenfalls der Nichtnegativitätsbedingung und – wegen der ganzzahligen Koeffizienten – ebenfalls der Ganzzahligkeitsbedingung unterliegen, kann man die Ungleichungen in Gleichungen umwandeln, die im Simplex-Tableau 1 dargestellt sind.

	x_1	x_2	
z	-7	1	0
x_3	16	-12	41
x_4	4	4	43
x_5	-4	4	-5

Simplex-Tableau 1: Ausgangslösung

Im Bild 1 ist das Problem graphisch dargestellt.

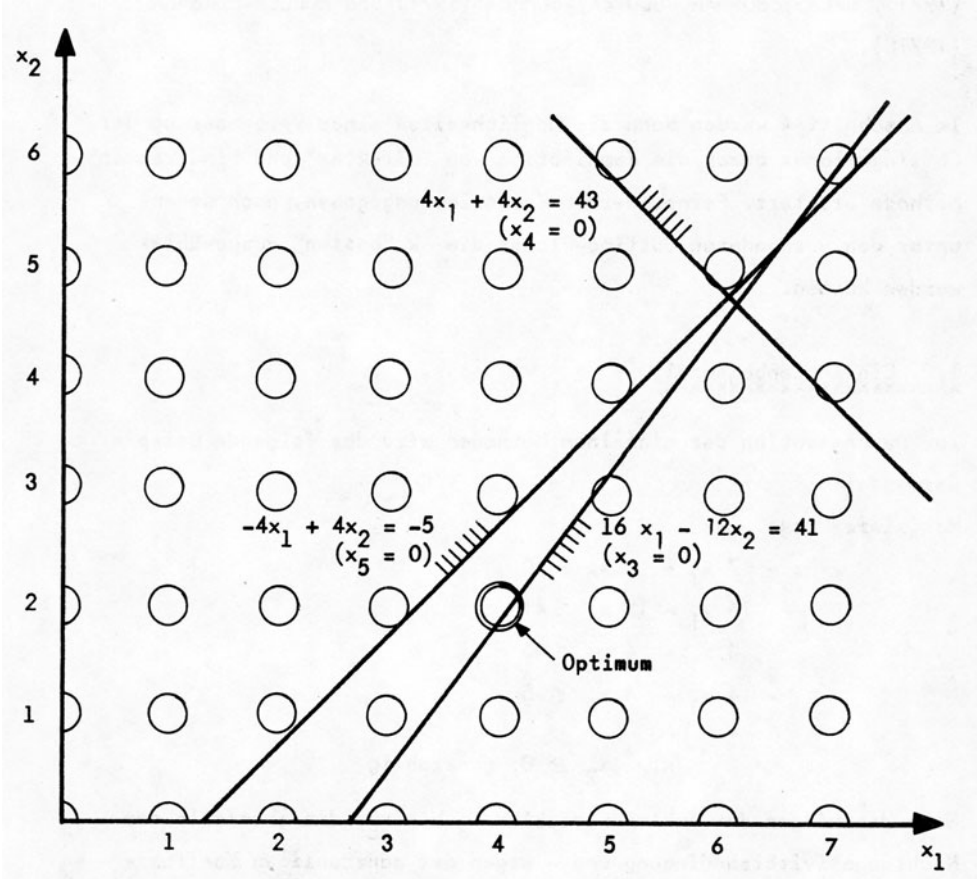


Bild 1: Graphische Darstellung des im Simplex-Tableau 1 genannten Optimierungsproblems

Das mit der Simplex-Methode gefundene nichtganzzahlige Optimum ist im Simplex-Tableau 2 gezeigt.

	x_3	x_4	
z	$\frac{2}{7}$	$\frac{17}{28}$	$\frac{1059}{28} = 37 \frac{23}{28}$
x_1	$\frac{1}{28}$	$\frac{3}{28}$	$\frac{85}{14} = 6 \frac{1}{14}$
x_2	$-\frac{1}{28}$	$\frac{1}{7}$	$\frac{131}{28} = 4 \frac{19}{28}$
x_5	$\frac{2}{7}$	$-\frac{1}{7}$	$\frac{4}{7}$

Simplex-Tableau 2: Nichtganzzahliges Optimum

Das ganzzahlige Optimum lautet: $x_1 = 4$, $x_2 = 2$, $x_3 = 1$,
 $x_4 = 19$, $x_5 = 3$, $z = 26$.

1.3. Notation xxxxxxxxxxxxxx

In dieser Arbeit wird folgende Notation verwendet:

$N_o = \{1, 2, \dots, n\}$ = Indexmenge der Nichtbasisvariablen der Ausgangs-
 lösung

$B_o = \{n+1, n+2, \dots, n+m\}$ = Indexmenge der Basisvariablen der
 Ausgangslösung; $N_o \cap B_o = \emptyset$

N = Indexmenge der Nichtbasisvariablen einer beliebigen Lösung,
 insbesondere einer Optimallösung

B = Indexmenge der Basisvariablen einer beliebigen Lösung, insbe-
 sondere einer Optimallösung

n = Anzahl der Nichtbasisvariablen

m = Anzahl der Restriktionen

c = Anzahl der Cutting-Planes eines Simplex-Tableaus

Die hier betrachteten Probleme lauten allgemein:

Maximiere z

$$z + \sum_{j \in N_0} a_{0j} x_j = a_{00}$$

$$x_{n+i} + \sum_{j \in N_0} a_{ij} x_j = a_{io} \quad \forall n+i \in B_0$$

$$x_j \geq 0 \quad \forall j \in \{B_0, N_0\}$$

$$x_j \text{ ganzzahlig} \quad \forall j \in N_0$$

$$a_{ij} \text{ ganzzahlig} \quad \forall i \in \{0, B_0\}, \quad \forall j \in \{0, N_0\}$$

und daher z ganzzahlig und

$$x_{n+i} \text{ ganzzahlig} \quad \forall n+i \in B_0$$

Die optimale nichtganzzahlige Lösung lautet:

$$z + \sum_{j \in N} \alpha_{0j} x_j = \alpha_{00}$$

$$x_k + \sum_{j \in N} \alpha_{ij} x_j = \alpha_{io} \quad \forall k \in B, \text{ wobei } x_k \text{ die Basisvariable der } i\text{-ten Restriktion ist}$$

$$x_j = 0 \quad \forall j \in N$$

Bei der Trennung ganzzahliger und nichtganzzahliger Teile eines Quotienten werden verwendet:

$$\left[\frac{a}{\lambda} \right] = \text{ganzzahliger Teil des Quotienten } \frac{a}{\lambda}; \quad \frac{a}{\lambda} - 1 < \left[\frac{a}{\lambda} \right] \leq \frac{a}{\lambda}$$

$$a^\lambda = \frac{a}{\lambda} - \left[\frac{a}{\lambda} \right] = \text{Positiver Divisionsrest; } 0 \leq a^\lambda < 1$$

Für $\lambda = 1$ ist $[a]$ der ganzzahlige Anteil von a ($a-1 < [a] \leq a$) und a^1 der gebrochene Teil von a ($0 \leq a^1 < 1$).

Mit GGT wird der größte gemeinsame Teiler bezeichnet.

2. Die "direkten" Cutting-Planes

2.1. Das Rechenprinzip der "ganzzahligen Division"

Als erstes soll nun das direkte Erzeugen von Cutting-Planes behandelt werden. Die direkten Cutting-Planes ergeben sich aus Gleichungen vom Typ $x_{n+i} + \sum_{j \in N} a_{ij} x_j = a_{io}$ durch "ganzzahlige Division" der Koeffizienten durch eine Konstante, die hier in Anlehnung an GOMORY (1963) mit λ bezeichnet werden soll. Dabei bedeutet "ganzzahlige Division", daß jeder nichtganzzahlige Quotient auf die nächstkleinere ganze Zahl abgerundet wird, z. B. $\left[\frac{5}{3} \right] = 1$, $\left[\frac{10}{13,7} \right] = 0$, $\left[\frac{27,9}{7,1} \right] = 3$, $\left[\frac{-4}{9} \right] = -1$, $\left[\frac{-27,9}{7,1} \right] = -4$. Aus der obigen Gleichung erhält man nach der Division durch λ die neue Gleichung:

$$x_{n+k} + \sum_{j \in N} a_{kj} x_j = a_{ko}$$

mit x_{n+k} (≥ 0 , ganzzahlig) als neuer Schlupfvariablen,

$$\text{und } a_{kj} = \left[\frac{a_{ij}}{\lambda} \right]$$

Daß diese neue Gleichung einen Cutting-Plane darstellt, der keine ganzzahligen Punkte abschneidet, ergibt sich aus der folgenden Überlegung. Die Ausgangsgleichung

$$x_{n+i} + \sum_{j \in N} a_{ij} x_j = a_{io}$$

kann man durch λ dividieren:

$$\frac{1}{\lambda} x_{n+i} + \sum_{j \in N} \frac{a_{ij}}{\lambda} x_j = \frac{a_{io}}{\lambda}$$

und in ganzzahlige und nichtganzzahlige Elemente trennen:

$$\left\{ 1^\lambda x_{n+i} + \sum_{j \in N} a_{ij}^\lambda x_j - a_{io}^\lambda \right\} + \sum_{j \in N} \frac{a_{ij}}{\lambda} x_j = \frac{a_{io}}{\lambda}$$

Der Ausdruck in der geschweiften Klammer muß stets größer als -1 sein, da $1^\lambda \geq 0$ und alle $a_{ij}^\lambda \geq 0$ sind und laut Definition $a_{io} < 1$ ist. Der Ausdruck in der geschweiften Klammer muß außerdem ganzzahlig sein, da - bei ganzzahligen x_j - die beiden anderen Ausdrücke der Gleichung ganzzahlig sind. Wegen dieser Ganzzahligkeit kann man $\{...\} > -1$ ersetzen durch $\{...\} \geq 0$, wodurch nur nichtganzzahlige Gebiete vom zulässigen Bereich abgeschnitten werden. Nach Einsetzen von $x_{n+k} = \{...\}$, $a_{kj} = \left\lceil \frac{a_{ij}}{\lambda} \right\rceil$ und $a_{ko} = \left\lceil \frac{a_{io}}{\lambda} \right\rceil$ erhält man den oben genannten Cutting-Plane als zusätzliche Bedingung:

$$x_{n+k} + \sum_{j \in N} a_{kj} x_j = a_{ko}$$

Unter Verzicht auf die jeweilige Schlupfvariable x_{n+k} erhält man die Cutting-Planes in Form von " \leq "-Bedingungen; diese Schreibweise wird in den Abschnitten 3.5 und 4 vorgezogen.

2.2. Beispiel

Am obigen Zahlenbeispiel sei das beschriebene Prinzip erläutert.
Aus der Gleichung

$$x_3 + 16 x_1 - 12 x_2 = 41$$

erhält man z. B. nach Division durch $\lambda = 3$

$$x_6 + 5 x_1 - 4 x_2 = 13,$$

nach Division durch $\lambda = \frac{16}{5} = 3,2$

$$x_7 + 5 x_1 - 4 x_2 = 12,$$

nach Division durch $\lambda = 4$

$$x_8 + 4 x_1 - 3 x_2 = 10,$$

nach Division durch $\lambda = 11$

$$x_9 + 1 x_1 - 2 x_2 = 3$$

und nach Division durch $\lambda = 16$

$$x_{10} + x_1 - x_2 = 2.$$

Aus der zweiten Gleichung

$$x_4 + 4x_1 + 4x_2 = 43$$

erhält man nach Division durch $\lambda = 4$

$$x_{11} + x_1 + x_2 = 10.$$

Und aus der dritten Gleichung

$$x_5 - 4x_1 + 4x_2 = -5$$

erhält man nach Division durch $\lambda = 2$

$$x_{12} - 2x_1 + 2x_2 = -3$$

und nach Division durch $\lambda = 4$

$$x_{13} - x_1 + x_2 = -2.$$

Durch eine graphische Darstellung der Cutting-Planes kann man leicht erkennen, welche Gebiete durch die einzelnen neuen Bedingungen abgeschnitten werden.

2.3. Die ungünstigsten Werte von λ

Als Divisionskonstanten λ eignen sich alle positiven Zahlen. Einige sind jedoch besonders gut geeignet. Bevor die Regeln zur Bestimmung der besonders geeigneten λ -Werte hergeleitet werden, ist in der Tabelle 1 gezeigt, welche verschiedenen Cutting-Planes aus der Bedingung $x_3 + 16x_1 - 12x_2 = 41$ bei Werten zwischen $\lambda \approx 2,93$ und $\lambda = 16$ entstehen.

Nr.	λ	Koeffizienten von		rechte Seite	Lösch- marke
		x_1	x_2		
1	$\frac{41}{3} + \varepsilon$ bis 16	1	-1	2	
2	12 bis $\frac{41}{3}$	1	-1	3	*
3	$\frac{41}{4} + \varepsilon$ bis $12 - \varepsilon$	1	-2	3	-
4	$\frac{41}{5} + \varepsilon$ bis $\frac{41}{4}$	1	-2	4	*
5	$8 + \varepsilon$ bis $\frac{41}{5}$	1	-2	5	*
6	$\frac{41}{6} + \varepsilon$ bis 8	2	-2	5	
7	6 bis $\frac{41}{6}$	2	-2	6	*
8	$\frac{41}{7} + \varepsilon$ bis $6 - \varepsilon$	2	-3	6	-
9	$\frac{16}{3} + \varepsilon$ bis $\frac{41}{7}$	2	-3	7	*
10	$\frac{41}{8} + \varepsilon$ bis $\frac{16}{3}$	3	-3	7	
11	$\frac{41}{9} + \varepsilon$ bis $\frac{41}{8}$	3	-3	8	*
12	$\frac{41}{10} + \varepsilon$ bis $\frac{41}{9}$	3	-3	9	*
13	$4 + \varepsilon$ bis $\frac{41}{10}$	3	-3	10	*
14	4	4	-3	10	
15	$\frac{41}{11} + \varepsilon$ bis $4 - \varepsilon$	4	-4	10	-
16	$\frac{41}{12} + \varepsilon$ bis $\frac{41}{11}$	4	-4	11	*
17	$\frac{16}{5} + \varepsilon$ bis $\frac{41}{12}$	4	-4	12	*
18	$\frac{41}{13} + \varepsilon$ bis $\frac{16}{5}$	5	-4	12	
19	3 bis $\frac{41}{13}$	5	-4	13	*
20	$\frac{41}{14} + \varepsilon$ bis $3 - \varepsilon$	5	-5	13	-

Tabelle 1: Cutting-Planes aus $x_3 + 16 x_1 = 12 x_2 = 41$ für verschiedene λ (ε = sehr kleine Zahl)

Die durch einen (*) markierten Cutting-Planes sind schwächer als die über ihnen stehenden, denn sie haben bei sonst gleichen Koeffizienten größere rechte Seiten. Ferner sind die mit einem Strich (-) gekennzeichneten Cutting-Planes ebenfalls schwächer als die über ihnen angegebenen, denn sie haben bei gleichen positiven Koeffizienten (und gleichen oder größeren rechten Seiten) absolut größere negative Koeffizienten. Interessant bleiben daher nur die Cutting-Planes Nr. 1, 6, 10, 14 und 18. Um diese, die sich in mindestens je einem positiven Koeffizienten unterscheiden, zu erzeugen, kann man wie folgt vorgehen.

Zunächst setzt man λ gleich dem größten positiven Koeffizienten der betrachteten Gleichung i , also

$$\lambda = \max_{j \in N} \{a_{ij}\},$$

und berechnet den ersten Cutting-Plane. Ein größeres λ würde zu keinem vernünftigen Cutting-Plane führen.

Dann verkleinert man λ schrittweise so weit, bis sich mindestens ein positiver Koeffizient des Cutting-Plane vergrößert. Das ist bei

$$\lambda_{\text{neu}} = \max_{j \in N} \left\{ \frac{a_{ij}}{\left\lfloor \frac{a_{ij}}{\lambda} \right\rfloor + 1} \mid a_{ij} > 0 \right\}$$

der Fall. Mit diesem $\lambda := \lambda_{\text{neu}}$ errechnet man den nächsten Cutting-Plane. Das wird fortgesetzt, bis λ so klein ist, daß kein guter Cutting-Plane mehr zu erwarten ist. Das ist im allgemeinen offensichtlich bei

$$\lambda = \text{GGT} \left\{ |a_{ij}| \right\}_{j \in N},$$

also spätestens bei $\lambda = 1$ der Fall.

Für das Beispiel der Tabelle beginnt man also mit $\lambda = 16$ (Cutting-Plane Nr. 1), berechnet dann $\lambda = \frac{16}{\left[\frac{16}{16}\right] + 1} = 8$ (Nr. 6),

dann $\lambda = \frac{16}{\left[\frac{16}{8}\right] + 1} = \frac{16}{3}$ (Nr. 10), dann $\lambda = \frac{16}{\left[\frac{16 \cdot 3}{16}\right] + 1} = 4$ (Nr. 14)

und kann nun wegen $\lambda = \text{GGT}\{16, 12\} = 4$ auf die Erzeugung weiterer Cutting-Planes verzichten.

Die Tabelle 1 hätte also nach Nr. 14 abgebrochen werden können. Tatsächlich ergibt sich der einzige noch nicht gestrichene spätere Cutting-Plane Nr. 18 auch aus der Summe von Nr. 1 und Nr. 14.

Wenn die Koeffizienten einer Restriktion (mit Ausnahme der rechten Seite) einen GGT $\neq 1$ haben, dann kann man diese Restriktion auch durch den Cutting-Plane ersetzen, der sich aus der ganzzahligen Division dieser Restriktion durch den GGT ergibt. Dadurch ändert sich dann nichts an der Restriktion, wenn die rechte Seite ebenfalls durch den GGT zu teilen ist; anderenfalls verläuft dieser Cutting-Plane parallel zu der ursprünglichen Restriktion, wobei der zulässige Bereich verkleinert wird. Im Beispiel könnte also die Bedingung

$$16 x_1 - 12 x_2 \leq 41$$

nach ganzzahliger Division durch $\lambda = \text{GGT} = 4$ ersetzt werden durch

$$4 x_1 - 3 x_2 \leq 10.$$

Von dieser neuen Restriktion könnten dann durch weitere ganzzahlige Division neue Cutting-Planes hergeleitet werden. Man würde damit im Beispiel die Cutting-Planes Nr. 1 und Nr. 6 der Tabelle 1 erhalten.

Aus gleichem Grund könnte die Bedingung

$$4 x_1 + 4 x_2 \leq 43$$

nach Division durch $\lambda = \text{GGT} = 4$ ersetzt werden durch

$$x_1 + x_2 \leq 10.$$

Entsprechend könnte die Bedingung

$$-4x_1 + 4x_2 \leq -5$$

nach Division durch ebenfalls $\lambda = \text{GGT} = 4$ ersetzt werden durch

$$-x_1 + x_2 \leq -2.$$

Immer wenn einzelne Bedingungen einen $\text{GGT} \neq 1$ aufweisen, dann ist ihr Ersatz durch die parallelen Cutting-Planes stets vorteilhaft.

2.4. Redundante Cutting-Planes

xx

Im allgemeinen sind viele Cutting-Planes redundant. Jedoch ist das oft nur mit erheblichem Aufwand zu prüfen.

Im Beispiel erkennt man leicht, daß die Cutting-Planes Nr. 6 und Nr. 10 schwächer sind als Nr. 1, da sie parallel verlaufen und relativ größere rechte Seiten haben. Man kann sie also ersatzlos streichen und behält nur noch die Cutting-Planes Nr. 1 und Nr. 14.

Bei größeren Problemen sind die Redundanzen meistens nicht so offensichtlich. So ist es im allgemeinen nicht zu empfehlen, eine vollständige Redundanzanalyse durchzuführen, da sie sehr aufwendig sein kann.

2.5. Die Restriktivität von Cutting-Planes

xx

Wenn auch die Redundanz von direkten Cutting-Planes schwer zu prüfen ist, so ist die Restriktivität von Cutting-Planes verhältnismäßig einfach zu messen. Man kann sie z. B. definieren durch den Abstand eines Cutting-Planes von dem nichtganzzahligen Optimum oder durch den Abstand eines Cutting-Planes vom ganzzahligen Optimum. Je weiter ein Cutting-Plane vom nichtganzzahligen Optimum entfernt

ist bzw. je näher er dem ganzzahligen Optimum ist, als desto restriktiver sei er bezeichnet.

Die Entfernung einer Bedingung oder eines Cutting-Planes

$x_{n+1} + \sum_{j \in N} a_{ij} x_j = a_{io}$ von einem Punkt $(x_j^0, j = 1, 2, \dots, n)$ ergibt sich bekanntermaßen aus der Formel

$$d = \frac{a_{io} - \sum_{j \in N} a_{ij} x_j^0}{\sqrt{\sum_{j \in N} (a_{ij})^2}}$$

Hier bedeutet ein negatives Vorzeichen von d , daß der Punkt auf der nicht zulässigen Seite des Cutting-Planes liegt.

Mit dieser Formel sollen nun von vier im Abschnitt 2.2 berechneten und nach den Darstellungen der Abschnitte 2.3 und 2.4 zu bestätigenden Cutting-Planes die Entfernungen d_{ng0} zu dem nichtganzzahligen Optimum $x_1 = 6 \frac{1}{14}$, $x_2 = 4 \frac{19}{28}$ und die Entfernungen d_{g0} zu dem ganzzahligen Optimum $x_1 = 4$, $x_2 = 2$ berechnet werden.

$$\text{Cutting-Plane } x_8 + 4 x_1 - 3 x_2 = 10$$

$$d_{ng0} = -0,05$$

$$d_{g0} = 0$$

$$\text{Cutting-Plane } x_{10} + x_1 - x_2 = 2$$

$$d_{ng0} = 0,43$$

$$d_{g0} = 0$$

$$\text{Cutting-Plane } x_{11} + x_1 + x_2 = 10$$

$$d_{ng0} = -0,53$$

$$d_{g0} = 2,83$$

$$\text{Cutting-Plane } x_{13} - x_1 + x_2 = -2$$

$$d_{ng0} = -0,43$$

$$d_{g0} = 0$$

Drei der vier Cutting-Planes schließen also das nichtganzzahlige Optimum aus (d_{ng0} negativ). Ferner gehen ebenfalls drei von ihnen genau durch das ganzzahlige Optimum ($d_{g0} = 0$).

Das nichtganzzahlige Optimum kann man mit der Simplex-Methode im allgemeinen rasch berechnen, so daß sich die d_{ng0} leicht bestimmen lassen.

Dagegen ist das ganzzahlige Optimum meistens nicht vor Ende der Rechnung mit Sicherheit bekannt. Jedoch kann man oft mit Näherungsmethoden eine gute ganzzahlige Lösung finden, die bei der Berechnung der Abstände an Stelle des ganzzahligen Optimums verwendet werden kann.

2.6. Direkte Cutting-Planes aus kombinierten Restriktionen

In den bisherigen Abschnitten wurden nur aus einzelnen Restriktionen Cutting-Planes entwickelt. Häufig ist es vorteilhaft, wenn man darüberhinaus auch aus konvexen Kombinationen der Bedingungen Cutting-Planes herleitet.

Beispielsweise kann man von den folgenden Bedingungen (vgl. Simplex-Tableau 1) ausgehen:

$$\begin{aligned}x_3 + 16 x_1 - 12 x_2 &= 41 \\x_4 + 4 x_1 + 4 x_2 &= 43 \\x_5 - 4 x_1 + 4 x_2 &= -5\end{aligned}$$

Aus der Kombination

$$(x_4 + x_5) + 0 x_1 + 8 x_2 = 38$$

erhält man mit $\lambda = 8$ den Cutting-Plane

$$x_{14} + 0 x_1 + x_2 = 4.$$

Aus der Kombination

$$(x_3 + 3 x_5) + 4 x_1 + 0 x_2 = 26$$

erhält man mit $\lambda = 4$ den Cutting-Plane

$$x_{15} + x_1 + 0 x_2 = 6 .$$

Aus der Kombination

$$(x_3 + 12 x_{14}) + 16 x_1 + 0 x_2 = 77$$

erhält man mit $\lambda = 16$ den wesentlich besseren Cutting-Plane

$$x_{16} + x_1 + 0 x_2 = 4 .$$

Aus der Kombination

$$(x_5 + 4 x_{16}) + 0 x_1 + 4 x_2 = 11$$

erhält man mit $\lambda = 4$ den Cutting-Plane

$$x_{17} + 0 x_1 + x_2 = 2 .$$

Die Restriktivität dieser einzelnen Cutting-Planes ließe sich wieder nach den in Abschnitt 2.5 angegebenen Regeln messen.

Bei großen Problemen ist es aufwendig, systematisch alle erfolgversprechenden Kombinationen von Bedingungen zu erproben. Vielmehr ist man bei der Wahl der Kombinationen auf Zufallsauswahlen und auf heuristische Regeln angewiesen (vgl. MÜLLER-MERBACH, 1970).

2.7. Die Verwendung direkter Cutting-Planes zur anfänglichen Reduktion

des zulässigen Bereichs

Abschließend zum Abschnitt Über die direkten Cutting-Planes sei in Anlehnung an zwei frühere Arbeiten (MÜLLER-MERBACH, 1970, Abschnitt IV, MÜLLER-MERBACH, 1971b, Abschnitt 7) die Einfügung von Cutting-Planes vor der Berechnung des ersten (nichtganzzahligen) Optimums vorgeschlagen.

Dabei werden zunächst nach den Regeln von Abschnitt 2.3 und evtl. auch 2.6 möglichst viele Cutting-Planes erzeugt. Sie können dann ggf. gemäß 2.4 auf Redundanz geprüft werden.

Falls die Zahl der Cutting-Planes sehr groß ist, empfiehlt es sich, die nach Abschnitt 2.5 am wenigsten restriktiven zu eliminieren. Dazu berechnet man erstens mit der Simplex-Methode (ohne Cutting-Planes) das nichtganzzahlige Optimum und wählt die Cutting-Planes mit den absolut größten negativen Abständen d_{ng0} ; und/oder zweitens berechnet man ganzzahlige Näherungslösungen und wählt die Cutting-Planes mit den kleinsten positiven Abständen d_{g0} .

Falls Cutting-Planes parallel zu ursprünglichen Restriktionen verlaufen, können diese Restriktionen gestrichen werden.

Nun wendet man auf das um die ausgewählten Cutting-Planes vergrößerte Problem die Simplex-Methode an und erhält im allgemeinen zwar wieder ein nichtganzzahliges Optimum, das aber dem ganzzahligen Optimum oft sehr viel näher ist als die ohne diese Cutting-Planes gefundene nichtganzzahlige Optimallösung. Um diese Lösung ganzzahlig zu machen, müssen nachträglich weitere Cutting-Planes eingefügt werden, und zwar nach der sog. "indirekten" Methode, wie sie im Abschnitt 3 behandelt wird.

Im obigen Beispiel erhält man mit den im Simplex-Tableau 3 eingefügten vier Cutting-Planes sofort das im Simplex-Tableau 4 dargestellte ganzzahlige Optimum. Da die Cutting-Planes mit x_8 , x_{11} und x_{13} parallel zu den Restriktionen von x_3 , x_4 und x_5 verlaufen, konnten diese gestrichen werden.

	x_1	x_2	
z	-7	1	0
x_8	4	-3	10
x_{10}	1	-1	2
x_{11}	1	1	10
x_{13}	-1	1	-2

Simplex-Tableau 3:
Ausgangslösung mit
vier Cutting-Planes

	x_8	x_{13}	
z	6	17	26
x_1	1	3	4
x_{10}	0	1	0
x_{11}	-2	-7	4
x_2	1	4	2

Simplex-Tableau 4:
Optimallösung mit
vier Cutting-Planes

Die in der Literatur vorgeschlagenen Methoden, die mit direkten Cutting-Planes arbeiten, erzeugen pro Simplex-Iteration genau einen direkten Cutting-Plane. Diese Methoden, die von GOMORY (1963), GLOVER (1968) und YOUNG (1968) etc. vorgeschlagen wurden, haben sich in der Praxis aber nicht bewährt. Aus diesem Grunde sollen hier die Methoden zur Erzeugung direkter Cutting-Planes im wesentlichen nur für die Einengung des zulässigen Bereichs vor Beginn der Berechnung des Optimums und darüberhinaus für die spätere Verschärfung von indirekten Cutting-Planes zur Verwendung vorgeschlagen werden.

3. Die "indirekt" erzeugten Cutting-Planes
oo

Im vorhergehenden Abschnitt 2 wurden die direkt aus den Nebenbedingungen herleitbaren Cutting-Planes behandelt. Demgegenüber sollen nun solche Cutting-Planes besprochen werden, die sich erst nach Berechnung des nichtganzzahligen Optimums bestimmen lassen. Sie werden hier - wie im Abschnitt 1 erwähnt - als "indirekt" bezeichnet,

Die Methoden, die mit diesen indirekten Cutting-Planes arbeiten, fügen in das jeweilige nichtganzzahlige Optimum einen Cutting-Plane ein, durch den dieses Optimum aus dem zulässigen Bereich ausgeschlossen wird. Mit Hilfe der dualen Simplex-Methode wird dann von diesem nicht mehr zulässigen Optimum aus ein neues zulässiges Optimum berechnet. Falls es die Ganzzahligkeitsbedingungen erfüllt, hat man das ganzzahlige Optimum gefunden; die Rechnung ist dann zu Ende. Anderenfalls bestimmt man an dem neuen Optimum wieder einen Cutting-Plane, der dieses Optimum ebenfalls aus dem zulässigen Bereich ausschließt, so daß man ein anderes zulässiges Optimum berechnen muß usw.

Die indirekten Cutting-Planes enthalten als Variablen nur einige oder alle Nichtbasisvariablen des jeweiligen Optimums ($j \in N$). Durch Einsetzen der entsprechenden Gleichungen der Ausgangslösung, was im Rechengang identisch mit einer Vektor-Matrix-Multiplikation ist, lassen sich diese Cutting-Planes auf die Nichtbasisvariablen der Ausgangslösung ($j \in N_0$) beziehen. Das ist zwar nicht wichtig für die bisherigen Methoden, die mit den indirekten Cutting-Planes arbeiten. Es erhöht aber die Anschaulichkeit der Cutting-Planes. Vor allem aber ist es wichtig für die Kombination von direkten und indirekten Methoden, die im Abschnitt 4 besprochen werden sollen. Aus diesem Grunde werden im folgenden alle indirekten Cutting-Planes auch auf die Nichtbasisvariablen der Ausgangslösung umgerechnet. - Bei der folgenden Illustration der einzelnen Methoden zur Erzeugung der indirekten Cutting-Planes wird das gleiche Beispiel wie bisher verwendet. Dabei wird von dem im Simplex-Tableau 2 angegebenen nichtganzzahligen Optimum ausgegangen.

3.1. Der DANTZIG-Cut

Der einfachste indirekte Cutting-Plane ist der von DANTZIG (1959). Er beruht auf der folgenden Überlegung. Wenn die vorhandene Optimallösung die Ganzzahligkeitsbedingungen nicht erfüllt, dann muß mit Sicherheit mindestens eine der gegenwärtigen Nichtbasisvariablen einen größeren Wert als Null annehmen. Wegen der auch für die Nichtbasisvariablen geltenden Ganzzahligkeitsbedingungen muß dieser Wert mindestens gleich Eins sein. Da man nicht weiß, welche Variable das sein wird, kann man fordern, daß die Summe aller Nichtbasisvariablen größer oder gleich Eins sein wird. Man erhält also die Bedingung

$$\sum_{j \in N} x_j \geq 1$$

oder

$$- \sum_{j \in N} x_j \leq -1$$

oder

$$x_{m+n+c} - \sum_{j \in N} x_j = -1 \quad \text{mit} \quad x_{m+n+c} \geq 0, \text{ ganzzahlig.}$$

Diese Bedingung ist der DANTZIG-Cut. Sie wird in die vorhandene Optimallösung als zusätzliche Nebenbedingung eingesetzt, wodurch die Lösung unzulässig wird. Mit Hilfe der dualen Simplex-Methode kann man nun eine neue Optimallösung berechnen.

Im Beispiel lautet der DANTZIG-Cut

$$x_3 + x_4 \geq 1$$

Dieser Cutting-Plane ist identisch mit der Bedingung

$$20 x_1 - 8 x_2 \leq 83,$$

was sich durch Einsetzen in die Ausgangslösung leicht bestätigen läßt.

3.2. Der erste GOMORY-Cut xxxxxxxxxxxxxxxxxxxxxxxxxxxx

Etwas stärker problemabhängig als der DANTZIG-Cut sind die verschiedenen GOMORY-Cuts, deren erster 1958 beschrieben wurde. Er geht aus von einer beliebigen Zeile der Optimallösung, deren rechte Seite nicht ganzzahlig ist. Diese Zeile sei

$$x_k + \sum_{j \in N} \alpha_{ij} x_j = \alpha_{io} \quad (x_k = \text{Basisvariable der } i\text{-ten Zeile; } k \in B).$$

Trennt man in dieser Zeile die ganzzahligen Teile der Koeffizienten $[\alpha_{ij}]$ von den nichtganzzahligen Teilen α_{ij}^1 ($0 \leq \alpha_{ij}^1 < 1$), so erhält man

$$\left\{ x_k + \sum_{j \in N} [\alpha_{ij}] x_j - [\alpha_{io}] \right\} + \sum_{j \in N} \alpha_{ij}^1 x_j = \alpha_{io}^1.$$

Der Ausdruck in der geschweiften Klammer muß wegen der ganzzahligen Koeffizienten bei ganzzahligen Variablen immer ganzzahlig sein. Ferner muß er stets kleiner als Eins sein, da α_{io}^1 stets kleiner als Eins ist und alle α_{ij}^1 mindestens gleich Null sind. Wegen der Ganzzahligkeit kann der Ausdruck in der geschweiften Klammer also maximal gleich Null sein. Ersetzt man diesen Ausdruck durch die negative Schlupfvariable $-x_{m+n+c}$ und multipliziert die ganze Gleichung mit minus Eins, so erhält man die Bedingung

$$x_{m+n+c} - \sum_{j \in N} \alpha_{ij}^1 x_j = -\alpha_{io}^1 \quad \text{mit } x_{m+n+c} \geq 0, \text{ ganzzahlig.}$$

oder ohne Schlupfvariable

$$- \sum_{j \in N} \alpha_{ij}^1 x_j \leq -\alpha_{io}^1$$

Diese Bedingung ist der GOMORY-Cut vom Typ "Method of Integer Forms" (GOMORY, 1958). Sie wird wie der DANTZIG-Cut in die vorhandene Optimallösung als zusätzliche Nebenbedingung eingefügt, wodurch die Lösung unzulässig wird. Eine neue Optimallösung erhält man mit der dualen Simplex-Methode.

Im Beispiel lassen sich aus der Zielfunktion sowie aus den drei Bedingungen die folgenden GOMORY-Cuts erstellen:

$$- \frac{2}{7} x_3 - \frac{17}{28} x_4 \leq - \frac{23}{28}$$

$$- \frac{1}{28} x_3 - \frac{3}{28} x_4 \leq - \frac{1}{14}$$

$$- \frac{27}{28} x_3 - \frac{1}{7} x_4 \leq - \frac{19}{28}$$

$$- \frac{2}{7} x_3 - \frac{6}{7} x_4 \leq - \frac{4}{7}$$

Durch Einsetzen der Ausgangsgleichungen erhält man diese Cutting-Planes in Abhängigkeit der Variablen x_1 und x_2 :

$$7 x_1 - x_2 \leq 37$$

$$x_1 \leq 6$$

$$16 x_1 - 11 x_2 \leq 45$$

$$8 x_1 \leq 48$$

3.3. Der zweite GOMORY-Cut xx

Auf einer ganz anderen Begründung als die bisher beschriebenen indirekten Cutting-Planes beruht der zweite GOMORY-Cut, der 1960 für gemischt-ganzzahlige Probleme entwickelt wurde. Wie BALAS (1971) erwähnt, ist dieser Cutting-Plane ein Spezialfall der Intersection Cuts (vgl. Abschnitt 3.4). Ferner wurde die gleiche Methode vom Verfasser (1971b), dem die Arbeit von GOMORY (1960) nicht zugänglich war, noch einmal entwickelt und unter dem Namen Two-Surface-Cut beschrieben.

Diese Methode geht wieder von einzelnen Zeilen der Optimallösung mit negativer rechter Seite aus, beispielsweise von

$$x_k + \sum_{j \in N} \alpha_{ij} x_j = \alpha_{i0} \quad (x_k = \text{Basisvariable der } i\text{-ten Zeile;} \\ k \in B).$$

Es wird nun gefragt, um welchen Betrag jedes x_j ($j \in N$) einzeln wachsen kann, bis x_k einen ganzzahligen Wert erreicht, der dem Wert α_{i0} benachbart ist. Für jedes positive α_{ij} ist das bei $x_j = \alpha_{i0} / \alpha_{ij}$ der Fall. Und für jedes negative α_{ij} ist das bei $x_j = (1 - \alpha_{i0}^1) / \alpha_{ij}$ erreicht. Wenn man diese Schnittpunkte verbindet, erhält man einen Cutting-Plane der folgenden Form:

$$- \sum_{j \in N} \alpha_{m+c,j} x_j \leq -1$$

oder mit der Schlupfvariablen x_{m+n+c}

$$x_{m+n+c} - \sum_{j \in N} \alpha_{m+c,j} x_j = -1 \quad \text{mit } x_{m+n+c} \geq 0.$$

Dabei gilt:

$$\alpha_{m+c,j} = \begin{cases} \alpha_{ij} / \alpha_{i0}^1, & \text{falls } \alpha_{ij} > 0 \\ 0, & \text{falls } \alpha_{ij} = 0 \\ \alpha_{ij} / (1 - \alpha_{i0}^1), & \text{falls } \alpha_{ij} < 0 \end{cases}$$

Es ist zu beachten, daß die neue Schlupfvariable x_{m+n+c} nicht generell der Ganzzahligkeitsbedingung unterliegt.

Wie vom Verfasser (1971b) gezeigt ist, führen die erste und die zweite Methode von GOMORY dann auf denselben Cutting-Plane, wenn für alle $j \in N$ die Bedingung $0 \leq \alpha_{ij} < 1$ gilt. In den anderen Fällen führt die zweite Methode meistens dann auf restriktivere Cutting-Planes, wenn negative α_{ij} vorhanden sind.

Im obigen Beispiel erhält man aus der Zielfunktion und den drei Restriktionen der Optimallösung die vier Cutting-Planes:

$$-\frac{8}{23}x_3 - \frac{17}{23}x_4 \leq -1$$

$$-\frac{1}{2}x_3 - \frac{3}{2}x_4 \leq -1$$

$$-\frac{1}{9}x_3 - \frac{4}{19}x_4 \leq -1$$

$$-\frac{1}{2}x_3 - \frac{1}{3}x_4 \leq -1$$

Durch Einsetzen der Ausgangsgleichungen erhält man diese Cutting-Planes wieder in Abhängigkeit der Variablen x_1 und x_2 :

$$\frac{196}{23}x_1 - \frac{28}{23}x_2 \leq \frac{1036}{23} \quad (\text{oder } 7x_1 - x_2 \leq 37)$$

$$14x_1 \leq 84 \quad (\text{oder } x_1 \leq 6)$$

$$\frac{448}{171}x_1 - \frac{84}{171}x_2 \leq \frac{2156}{171} \quad (\text{oder } 16x_1 - 3x_2 \leq 77)$$

$$\frac{28}{3}x_1 - \frac{14}{3}x_2 \leq \frac{203}{6} \quad (\text{oder } 8x_1 - 4x_2 \leq 29)$$

3.4. Intersection Cuts xxxxxxxxxxxxxxxxxxxxxxxxxxxx

Eine neue Art von Cutting-Planes ist von BALAS (1971) beschrieben worden, nämlich die der Intersection Cuts. Hier wird durch die ganzzahligen Punkte, die das nichtganzzahlige Optimum umgeben, eine konvexe Hyperfläche gelegt. Dann werden die Schnittpunkte, die diese Fläche mit den im Optimum sich schneidenden Kanten bildet, berechnet. Die Hyperebene, die durch diese Schnittpunkte geht, ist der Cutting-Plane. BALAS (1971) schlug als Hyperfläche die Hyperkugel (bzw. den Kreis im zwei-, die Kugel im dreidimensionalen Raum) vor. Von

BALAS, BOWMAN, GLOVER, SOMMER (1971) wird dagegen ein "Hyperoktaeder" (bzw. das Quadrat im zwei-, das Oktaeder im dreidimensionalen Raum) bevorzugt. Man kann ferner Ellipsoide und andere konvexe Hyperflächen verwenden. Ein Extremfall des Ellipsoids ist das aus zwei parallelen Ebenen bestehende Gebilde, das den im Abschnitt 3.3 behandelten Cutting-Planes zugrunde liegt.

Die formelmäßige Herleitung der Intersection Cuts ist etwas aufwendig, so daß hier darauf verzichtet wird.

Im Beispiel wurden die zu bildenden Hyperflächen durch die Punkte (x_1, x_2) gleich $(6,4)$, $(6,5)$, $7,5$ und $(7,4)$ gehen. Wenn man nach BALAS, BOWMAN, GLOVER und SOMMER (1971) ein Quadrat wählt, so besteht es aus den Seiten $x_1 + x_2 = 10$, $x_1 - x_2 = 1$, $x_1 + x_2 = 12$ und $x_1 - x_2 = 3$. Die Schnittpunkte mit den sich im Optimum schneidenden Kanten $16 x_1 - 12 x_2 = 41$ und $4 x_1 + 4 x_2 = 43$ haben die Koordinaten $(5 \frac{3}{4}, 4 \frac{1}{4})$ und $(5 \frac{7}{8}, 4 \frac{7}{8})$.

Die Verbindung beider Punkte führt zu dem Cutting-Plane

$$-\frac{2}{11} x_3 - \frac{1}{3} x_4 \leq -1$$

Durch Einsetzen der Ausgangsgleichungen läßt sich diese Bedingung auf x_1 und x_2 beziehen:

$$\frac{140}{33} x_1 - \frac{28}{33} x_2 \leq \frac{686}{33} \quad (\text{oder } 10 x_1 - 2 x_2 \leq 49).$$

3.5. Modifikationen von indirekten Cutting-Planes

xx

Die meisten errechneten indirekten Cutting-Planes lassen sich nach der im Abschnitt 2 beschriebenen Methode, die letztlich auf GOMORY (1963) zurückgeht, modifizieren, worauf auch BALAS (1971) S. 27-28 hinweist. Allerdings ist nicht sicher, daß man damit die Cutting-Planes immer restriktiver macht.

Beispielsweise kann man den im Abschnitt 3.3 berechneten Cutting-Plane

$$-\frac{1}{9}x_3 - \frac{4}{19}x_4 \leq -1$$

ganzzahlig durch $\lambda = \frac{1}{9}$ teilen und erhält den neuen Cutting-Plane

$$-x_3 - 2x_4 \leq -9 \quad (\text{oder } 24x_1 - 4x_2 \leq 118).$$

Derartige Modifikationen kann man bei allen Cutting-Planes versuchen, und zwar mit verschiedenen Werten von λ . Die günstigsten Werte von λ kann man nach ähnlichen Beziehungen wie im Abschnitt 2.3 bestimmen. Man beginnt mit

$$\lambda = \max_{j \in N} \{ -\alpha_{ij} \}$$

und verkleinert λ schrittweise nach der Beziehung

$$\lambda_{\text{neu}} = \max_{j \in N} \left\{ \frac{-\alpha_{ij}}{\left[\frac{-\alpha_{ij}}{\lambda} \right] + 1} \mid \alpha_{ij} < 0 \right\},$$

wobei dann $\lambda := \lambda_{\text{neu}}$ gesetzt wird. Mit jedem λ wird ein neuer Cutting-Plane berechnet. Durch Redundanzprüfungen wie im Abschnitt 2.4 und Restriktivitätsberechnungen wie im Abschnitt 2.5 kann man die Überflüssigen und wenig restriktiven Cutting-Planes eliminieren.

4. Die Kombination von direkten und indirekten Methoden zur Erzeugung von Cutting-Planes

Die bisher Überwiegend angewendeten Cutting-Plane-Methoden fügen in jedem Schritt einen einzigen indirekten Cutting-Plane in die Optimallösung ein und wenden dann die duale Simplex-Methode an, bis ein neues Optimum gefunden ist (vgl. Anfang des Abschnittes 3).

$$\frac{140}{33} x_1 - \frac{28}{33} x_2 \leq \frac{686}{33} .$$

Sein Abstand vom ganzzahligen Optimum beträgt $d_{g0} = 1,67$.

Der Abstand vom nichtganzzahligen ersten Optimum lautet $d_{ng0} = -0,23$ (vgl. Abschnitt 2.5). Teilt man den Cutting-Plane in der auf die Ausgangskoordinaten bezogenen Form ganzzahlig durch $\lambda = \frac{28}{33}$, so erhält man den neuen Cutting-Plane

$$5 x_1 - x_2 \leq 24 .$$

Die Abstände dieses Cutting-Planes von den Optima betragen

$d_{g0} = 1,18$ bzw. $d_{ng0} = -0,33$. Umgerechnet auf die Variablen x_3 und x_4 lautet der gleiche Cutting-Plane

$$-\frac{6}{28} x_3 - \frac{11}{28} x_4 \leq -1 \frac{19}{28} .$$

Teilt man diese Bedingung ganzzahlig durch $\lambda = \frac{6}{28}$, so erhält man den neuen Cutting-Plane

$$-x_3 - 2 x_4 \leq -8$$

bzw. $24 x_1 - 4 x_2 \leq 119$.

Die Abstände von den Optima betragen $d_{g0} = 1,60$ bzw. $d_{ng0} = -0,33$.

Teilt man diese Bedingung nun ganzzahlig durch $\lambda = 4$, so ergibt sich der weitere Cutting-Plane

$$6 x_1 - x_2 \leq 29$$

bzw. $-\frac{1}{4} x_3 - \frac{1}{2} x_4 \leq -5 \frac{3}{4}$.

Dieser Cutting-Plane ist von den Optima $d_{g0} = 1,15$ bzw.

$d_{ng0} = -0,45$ entfernt.

Wenn man die Entfernungen d_{ng0} der einzelnen Cutting-Planes vom nichtganzzahligen Optimum vergleicht, so sieht man deutlich die Verschärfung der Cutting-Planes von $d_{ng0} = -0,23$ über $-0,33$

(zweimal) bis $-0,45$. Durch Anwendung von direkten Methoden der Cutting-Plane-Erzeugung auf einen indirekten Cutting-Plane wurde hier der Abstand vom nichtganzzahligen Optimum also fast verdoppelt.

In diesem Beispiel haben sich die d_{ng0} vergrößert, während sich die d_{g0} nicht so signifikant änderten. Andere Beispiele verhalten sich entgegengesetzt. So hat der im Abschnitt 3.2 berechnete Cutting-Plane

$$-\frac{27}{28}x_3 - \frac{1}{7}x_4 \leq -\frac{19}{28}$$

bzw.
$$16x_1 - 11x_2 \leq 45$$

die Abstände $d_{g0} = 0,15$ bzw. $d_{ng0} = -0,04$. Teilt man die auf x_1 und x_2 bezogene Bedingung ganzzahlig durch $\lambda = 16$, so erhält man den neuen Cutting-Plane

$$x_1 - x_2 \leq 2$$

bzw.
$$-\frac{1}{14}x_3 + \frac{1}{28}x_4 \leq \frac{17}{28}$$

mit den Abständen $d_{g0} = 0$ bzw. $d_{ng0} = 0,43$. Dieser Cutting-Plane schließt zwar das gegenwärtige nichtganzzahlige Optimum nicht aus dem zulässigen Bereich aus, berührt aber dafür bereits das ganzzahlige Optimum.

Die Zahl der verschiedenen Cutting-Planes, die sich mit den hier skizzierten Methoden aus einem einzigen direkten Cutting-Plane erzeugen lassen, kann sehr groß sein. Und zwar kann man aus den sowohl auf die Variablen x_j ($j \in N$) als auch auf die Variablen x_j ($j \in N_0$) bezogenen Cutting-Planes durch ganzzahliges Teilen neue Cutting-Planes erzeugen. Die Qualität, d. h. Restriktivität der einzelnen Cutting-Planes kann man am Abstand d_{g0} zum ganzzahligen

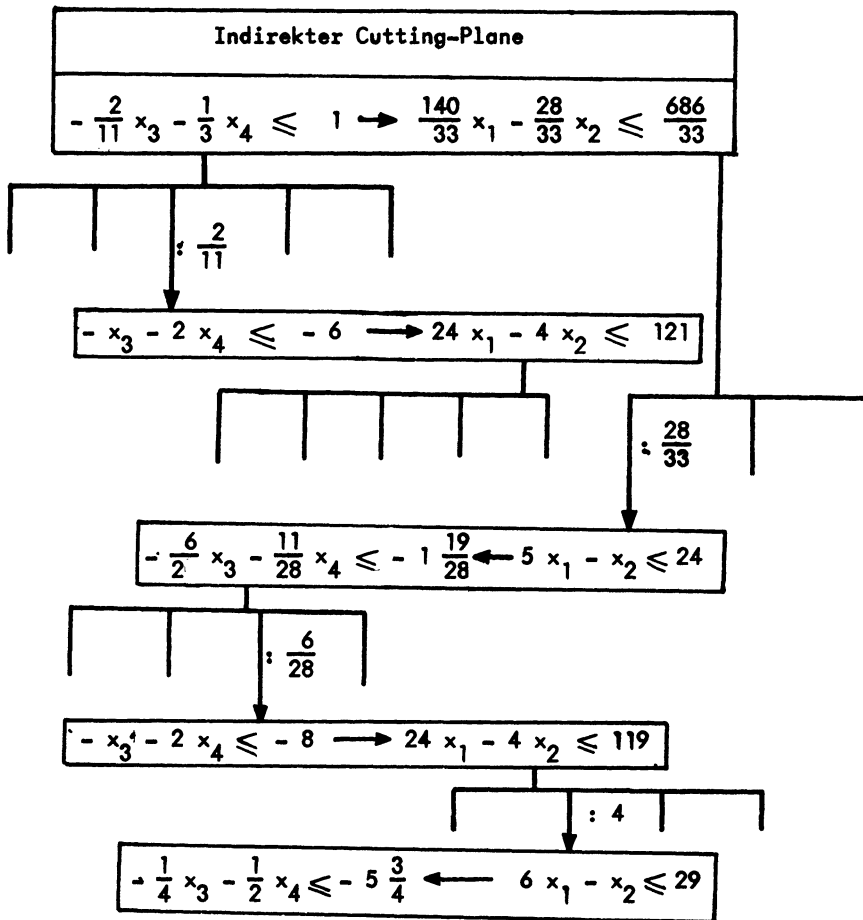


Bild 2: Baum der aus einem indirekten Cutting-Plane erzeugbaren Cutting-Planes

Je größer der negative Abstand zum nichtganzzahligen Optimum, je kleiner der positive Abstand zur besten bekannten ganzzahligen Lösung (der nicht negativ sein kann) und je größer die Verringerung des Zielfunktionswertes ist, desto geeigneter ist ein Cutting-Plane. Wie im einzelnen diese drei Kriterien zu gewichten sind, muß noch in Testreihen erprobt werden.

In der Tabelle 2 sind 18 verschiedene Cutting-Planes, die analog zum Baum von Bild 2 hergeleitet wurden, und ihre Herkunftsregeln angegeben. Ihre Auswahl ist recht willkürlich. Nur einige gehören zu den restriktivsten Cutting-Planes, insbesondere die mit den Basisvariablen x_9 , x_{14} bis x_{19} und x_{22} .

Eine Sonderstellung nimmt der Cutting-Plane von x_{22} ein. Er ergibt sich aus der Zielfunktion, die um das größte Δz , hier also um 6,08 von x_{14} , gegenüber dem nichtganzzahligen Optimum ganzzahlig verschoben wird. Die Begründung ist einfach. Der Cutting-Plane von x_{14} gibt an, daß der Zielfunktionswert des ganzzahligen Optimums um mindestens $\Delta z = 6,08$ unter dem Zielfunktionswert des derzeitigen nichtganzzahligen Optimums liegen muß. Wegen der ganzzahligen Ausgangskoeffizienten der Zielfunktion muß aber auch z im ganzzahligen Optimum ganzzahlig sein. Daraus folgt, daß $z \leq 31$ sein muß. Also gilt $\frac{2}{7} x_3 + \frac{17}{28} x_4 \geq 6 \frac{23}{28}$.

4.3. Die duale Simplex-Methode mit mehreren Cutting-Planes xxx

Wenn die restriktivsten Cutting-Planes gewählt sind, kann man mit der dualen Simplex-Methode ein neues zulässiges, aber im allgemeinen auch nur nichtganzzahliges Optimum finden, von wo aus man wieder neue Cutting-Planes bestimmt.

Das gleichzeitige Einfügen von mehreren Cutting-Planes (vor allem der restriktivsten) hat den Vorteil, daß man sich dem ganzzahligen Optimum schneller nähert als bei einem einzigen Cutting-Plane.

Im Beispiel kann man etwa die Cutting-Planes mit den Basisvariablen x_{10} ($d_{g0} = 0$), x_{14} (absolut größtes d_{ng0}) und x_{22} (größtes Δz) in das Simplex-Tableau 2 der Optimallösung aufnehmen. Als neues Optimum erhält man dann mit der dualen Simplex-Methode z. B. die Lösung des Simplex-Tableaus 5.

	x_5	x_{22}	
z	0	1	31
x_1	$\frac{1}{24}$	$\frac{1}{6}$	$4 \frac{23}{24}$
x_2	$\frac{7}{24}$	$\frac{1}{6}$	$3 \frac{17}{24}$
x_3	$\frac{17}{6}$	$-\frac{2}{3}$	$6 \frac{1}{6}$
x_4	$-\frac{4}{3}$	$-\frac{4}{3}$	$8 \frac{1}{3}$
x_{10}	$\frac{1}{4}$	0	$\frac{3}{4}$
x_{14}	$-\frac{1}{24}$	$-\frac{3}{4}$	$\frac{1}{24}$

Simplex-Tableau 5: Nichtganzzahliges Optimum
nach Hinzufügung von 3
Cutting-Planes

Ein weiterer Cutting-Plane, nämlich

$$-\frac{1}{24} x_3 - \frac{1}{6} x_4 \leq -\frac{23}{24}$$

oder

$$x_1 \leq 4 ,$$

der sich nach GOMORY aus der Zeile von x_1 herleiten läßt, genügt zum Auffinden des ganzzahligen Optimums. Allerdings wird man auch hier zunächst mehrere Cutting-Planes bestimmen und die restriktivsten in das Simplex-Tableau einfügen.

Wenn wie im Simplex-Tableau 5 die Zielfunktion Nullen enthält, gibt es offenbar mehrere gleich gute Lösungen. In diesen Fällen scheint es für die Restriktivität der neu zu erzeugenden Cutting-Planes vorteilhaft zu sein, diejenige der gleich guten Lösungen zu wählen, die vom ganzzahligen Optimum (bzw. der besten bekannten ganzzahligen Lösung) am weitesten entfernt ist. Am obigen Beispiel läßt sich das deutlich zeigen. Hier sei jedoch auf Einzelheiten verzichtet.

4.4. Ein Verfahrensvorschlag xx

Abschließend sei eine Verfahrenskombination zur Lösung von "Pure-Integer"-Problemen vorgeschlagen.

1. Bestimmung einer guten ganzzahligen Lösung durch eine oder mehrere Näherungsmethoden.
2. Berechnung des nichtganzzahligen Optimums mit der Simplex-Methode.
3. Bestimmung von direkten Cutting-Planes aus den Ausgangsrestriktionen (gemäß Abschnitt 2.1 bis 2.4 und 2.6) und Einfügen der restriktivsten von ihnen (nach Abschnitt 2.5) in das Ausgangstableau.
4. Erneute Berechnung eines (nichtganzzahligen) Optimums, aber jetzt für das um die Cutting-Planes von Schritt 3 modifizierte Problem.

5. Falls die vorliegende Lösung die Ganzzahligkeitsbedingung erfüllt,
→ Ende. Sonst → Schritt 6.
6. Löschen derjenigen Cutting-Planes, die weder das gegenwärtige nichtganzzahlige Optimum eingrenzen noch der im Schritt 1 bestimmten ganzzahligen Lösung nahe sind. (Dieser Schritt dient nur dazu, das Simplex-Tableau klein zu halten.)
7. Bestimmung von indirekten und (von ihnen abgeleiteten) direkten Cutting-Planes (gemäß Abschnitt 4.1) und Einfügen der restriktivsten von ihnen (nach Abschnitt 4.2) in das Simplex-Tableau der gegenwärtigen Optimallösung. Insbesondere sollte ein der Zielfunktion paralleler Cutting-Plane eingefügt werden, der um mindestens das maximale Δz vom gegenwärtigen Zielfunktionswert entfernt ist (vgl. x_{22} im Abschnitt 4.2).
8. Berechnung eines neuen Optimums mit der dualen Simplex-Methode und → Schritt 5.

Eventuell könnten die Schritte 3 und 4 mehrfach hintereinander durchlaufen werden.

Die hier beschriebene Verfahrenskombination wurde bisher nur an kleinen Beispielen getestet. Die Ergebnisse sind ermutigend. Ein Test an größeren Problemen ist vorgesehen.

Der mit der ganzzahligen Optimierung noch weniger vertraute Leser sei auf die zusammenfassenden Veröffentlichungen von BALINSKI (1965), GEOFFRION und MARSTEN (1972) sowie KORBUT und FINKELSTEIN (1971) zur Einarbeitung in dieses Gebiet hingewiesen.

5. Literatur

ooooo
Egon BALAS (1965): An Additive Algorithm for Solving Linear Programs with Zero-One Variables. Operations Research 13 (1965) 4, S.517 - 549

Egon BALAS (1971): Intersection Cuts - A New Type of Cutting Planes for Integer Programming. Operations Research 19 (1971) 1, S. 19 - 39

Egon BALAS, V. Joseph BOWMAN, Fred GLOVER, David SOMMER (1971): An Intersection Cut from the Dual of the Unit Hypercube. Operations Research 19 (1971) 1, S. 40 - 44

M. L. BALINSKI (1965): Integer Programming: Methods, Uses, Computation. Management Science 12 (1965) 3, S. 253 - 313

George B. DANTZIG (1959): Note on Solving Linear Programs in Integers. Naval Research Logistics Quarterly 4 (1959) 1, S. 75 - 76

R. J. DAKIN (1963): A Tree-Search Algorithm for Mixed Integer Programming Problems. The Computer Journal 8 (1963) 3, S. 250 - 255

R. E. ECHOLS, L. COOPER (1968): Solution of Integer Linear Programming Problems by Direct Search. Journal of the ACM 15 (1968), S. 75 - 84

A. M. GEOFFRION, R. E. MARSTEN (1972): Integer Programming Algorithms: A Framework and State-of-the-Art Survey. Management Science 18 (1972) 9, S. 465 - 491

Fred GLOVER (1968): A New Foundation for a Simplified Primal Integer Programming Algorithm. Operations Research 16 (1968) 4, S. 727 - 740

Ralph E. GOMORY (1958): An Algorithm for Integer Solutions to Linear Programs. Princeton-IBM Research Project Technical Report 1958, nachgedruckt in: Recent Advances in Mathematical Programming, hrsg. von R. L. GRAVES und Ph. WOLFE, New York 1963, S. 269 - 302

Ralph E. GOMORY (1960): An Algorithm for the Mixed Integer Problem. Rand Report RM-2597, The Rand Corporation 1960

Ralph E. GOMORY (1963): An All-Integer Integer Programming Algorithm. Industrial Scheduling, hrsg. von J. F. MUTH und G. L. THOMPSON, Englewood Cliffs 1963, S. 193 - 206

Frederick S. HILLIER (1969): Efficient Heuristic Procedures for Integer Linear Programming with an Interior. Operations Research 17 (1969) 4, S. 600 - 637

A. A. KORBUT, J. J. FINKELSTEIN (1971): Diskrete Optimierung (deutsche Übersetzung), Berlin 1971

Heinz KREUZBERGER (1969a): Ein Näherungsverfahren zur Bestimmung ganzzahliger Lösungen bei linearen Optimierungsproblemen. Ablauf- und Planungsforschung 9 (1969) 3, S. 137 - 152

Heinz KREUZBERGER (1969b): Verfahren zur Lösung ganzzahliger linear Optimierungsprobleme (Dissertation). Darmstadt 1969

Heinz KREUZBERGER (1970): Numerische Erfahrungen mit einem heuristischen Verfahren zur Lösung ganzzahliger linearer Optimierungsprobleme. Elektronische Datenverarbeitung 12 (1970) 7, S. 289 - 306

A. H. LAND, A. G. DOIG (1960): An Automatic Method of Solving Discrete Programming Problems. Econometrica 28 (1960) 3, S. 497 - 519

Heiner MÜLLER-MERBACH (1969): Das Verfahren der "vorsichtigen Annäherung" - Eine heuristische Methode zur Lösung gewisser Probleme der ganzzahligen Planungsrechnung. Elektronische Datenverarbeitung 11 (1969) 12, S. 464 - 466

Heiner MÜLLER-MERBACH (1970): Approximation Methods for Integer Programming. Forschungsbericht Nr. 26-1970 des Lehrstuhls für Betriebswirtschaftslehre an der Johannes Gutenberg-Universität, Mainz 1970

Heiner MÜLLER-MERBACH (1971a): Operations Research - Methoden und Modelle der Optimalplanung (2. Aufl.), München 1971

Heiner MÜLLER-MERBACH (1971b): The Two-Surface-Cut - A Contribution to Cutting Plane Techniques in Integer Linear Programming. Forschungsbericht Nr. 30-1971 des Lehrstuhls für Betriebswirtschaftslehre an der Johannes Gutenberg-Universität, Mainz 1971

R. D. YOUNG (1968): A Simplified Primal (All-Integer) Integer Programming Algorithm. Operations Research 16 (1968) 4, S. 750 - 782

AUTORENVERZEICHNIS

- BECKER, Ursel, Dipl.-Math., Posttechnisches Zentralamt, Referat für Statistik und Ökonometrie, 61 Darmstadt, Wilhelminenstrasse 1-3
- BOL, Georg, Dipl.-Math., Institut für Statistik und Quantitative Methoden der Unternehmensführung der Universität Karlsruhe, 75 Karlsruhe, Postfach 6380
- BRAUERS, W. K., Dr., Center for Research in Defense, Department for Operations Research, 2de Regiment Lansierslaan 1, B-1040 Brüssel
- BÜHLER, Wolfgang, Dipl.-Math., Lehrstuhl für Unternehmensforschung, RWTH Aachen, 51 Aachen, Templergraben 55
- DATHE, H. M., Dr., IABG, 8012 Ottobrunn, Robert-Koch-Strasse Geb. 23
- DIRICKX, Ivo M. I., Dr., International Institute of Management, 1 Berlin 33, Griegstrasse 5
- DRNAS, Thomas M., PhD, Head, Reliability Section, Research and Development Division, Hughes Aircraft Co., Culver City, California, USA
- DÜRR, Walter, Dipl.-Math., Universität Regensburg, Fachbereich Wirtschaftswissenschaft, 84 Regensburg, Postfach 397
- ECKHARDT, Ulrich, Dr., Zentralinstitut für angewandte Mathematik, Kernforschungsanlage Jülich GmbH, 517 Jülich, Postf. 365
- EMRICH, Ortwin, Dr., Lehrstuhl für Statistik, Universität Augsburg, 89 Augsburg, Memminger Str. 6/14
- GAL, Tomas, Prof. Dr., Lehrstuhl für Unternehmensforschung, RWTH Aachen, 51 Aachen, Templergraben 55
- GOLDSTEIN, Bernd H., Dr., Institut für Statistik der Universität Karlsruhe, 75 Karlsruhe, Postfach 6380
- GORR, W. L., Institute of Physical Planning, Carnegie-Mellon University, Schenley Park, Pittsburgh, Pennsylvania 15213, USA
- HAEHLING von LANZENAUER, Christoph, Prof. Dr., School of Business Administration, University of Western Ontario, London 72/Canada
- HENKE, Manfred, Dr., Deutsche Gesellschaft für Unternehmensforschung, 53 Bonn, Lennéstrasse 37
- HENN, Rudolf, Prof. Dr., Institut für Statistik der Universität Karlsruhe, 75 Karlsruhe 1, Postfach 6380
- HOCHSTÄDTER, Dieter, Dr., Institut für Angewandte Mathematik der Technischen Hochschule München, 8 München 2, Arcisstrasse 21
- JAEGER, Arno, Prof. Dr., Seminar für Theoretische Wirtschaftslehre, Ruhr-Universität, 463 Bochum
- JUNGINGER, Werner, Dr., Institut für Informatik und Universitätsrechenzentrum, Universität Stuttgart, 7 Stuttgart 1, Herdweg 51
- KORTANEK, K. O., Prof. Dr., Institute of Physical Planning, Carnegie-Mellon University, Schenley Park, Pittsburgh, Pennsylvania 15213, USA
- KOSTEN, L., Prof. Dr., Technische Hogeschool Delft, Delft-8, Juliana-laan 132, Holland

- KÜNZI, Hans, Prof. Dr., Regierungsrat, Institut für Operations Research, Universität Zürich, CH 8006 Zürich, Weinbergstr. 59
- LIESENFELD, K. P., Dr., Lehrstuhl für Unternehmensforschung, RWTH Aachen, 51 Aachen, Templergraben 55
- LIN, Chi-Yuan, Prof. Dr., Department of Quantitative Business Analysis, University of Southern California, University Park, Los Angeles, California 90007, USA
- MEISE, Jörg, Dipl.-Ing., Battelle-Institut e. V., 6 Frankfurt 90, Postfach 900 160
- MENSCH, Gerhard, Prof. Dr., International Institute of Management, 1 Berlin 33, Griegstr. 5
- MEYHAK, Hermann, Dr., Industrieseminar der Universität Mannheim (WH), 68 Mannheim, Schloss
- MOESCHLIN, Otto, Dr., Institut für Statistik der Universität Karlsruhe, 75 Karlsruhe 1, Postfach 6380
- MÜLLER-MERBACH, Heiner, Prof. Dr., Technische Hochschule Darmstadt, Fachgebiet Betriebswirtschaftslehre, 61 Darmstadt, Hochschulstr. 1
- OPITZ, Otto, Prof. Dr., Lehrkanzel für Nationalökonomie und Sozialpolitik, Universität Innsbruck, A-6020 Innsbruck, Adolf-Pichler-Platz 6/I
- ORDELHEIDE, Dieter, Dr., Institut für Unternehmensführung und Unternehmensforschung, Ruhr-Universität, 463 Bochum
- PETERS, Lutz, Dr., McKinsey & Company, Inc., 4 Düsseldorf, Jägerhofstrasse 12
- RABUSSEAU, R., Dipl.-Kfm., Institut für Unternehmensführung und Unternehmensforschung, Sektion Mathematik und Informatik, Ruhr-Universität, 463 Bochum
- REICH, W., cand. math., Universität Heidelberg
- RÖDDER, W., Dipl.-Math., Lehrstuhl für Unternehmensforschung, RWTH Aachen, 51 Aachen, Templergraben 55
- RUTSCH, Martin, Prof. Dr., Institut für Statistik und Quantitative Methoden der Unternehmensführung der Universität Karlsruhe, 75 Karlsruhe 1, Postfach 6380
- SCHICK, George J., Prof. Dr., Chairman, Quantitative Analysis, University of Southern California, University Park, Los Angeles, California 90007, USA
- SCHIPS, Bernd, Prof. Dr., Seminar für Theoretische Wirtschaftslehre, Ruhr-Universität, 463 Bochum
- SCHNEEWEISS, Christoph, Prof. Dr., Lehrstuhl für Operations Research, Freie Universität Berlin, 1 Berlin 33, Thielallee 66
- SEILER, Karl, III, 1229 Old Stable Road, McLean, Virginia 22101, USA
- SIERENBERG, R. W., Technische Hogeschool Delft, Delft-8, Julianalaan 132, Holland
- STECKHAN, Helmut, Dr., Uni.-Doz., Alfred-Weber-Institut, Universität Heidelberg, 69 Heidelberg, Bergheimer Str. 104
- STIER, Winfried, Dr., Seminar für Theoretische Wirtschaftslehre, Ruhr-Universität, 463 Bochum

SUTTON, A. M. , Operational Research Department, Atkins Planning,
Woodcote Grove, Ashley Road, Epsom, Surrey, England

TAWADROS, Milad A. , Prof. Dr. , Indiana University, 1825 Northside
Boulevard, South Bend, Indiana 46615, USA

UEBE, Götz, Dr. , Institut für Angewandte Mathematik, Technische Uni-
versität München, 8 München 2, Arcisstr. 21

WEGENER, Michael, Dipl.-Ing. , Batelle-Institut e. V. , 6 Frankfurt 90,
Postfach 900 160

physica paperback

Bamberg, Günter

Statistische Entscheidungstheorie

1972. ca. 130 S. DM 20,— ISBN 3 7908 0099 6

Basler, Herbert

Grundbegriffe der Wahrscheinlichkeitsrechnung und statistischen Methodenlehre

Mit 27 Beispielen u. 35 Aufgaben mit Lösungen.

3. Aufl. 1971. 147 Seiten. DM 14,—

ISBN 3 7908 0098 8

Berg, Claus C.

Programmieren mit FORTRAN

in Vorb. ISBN 3 7908 0120 8

Bliefernich-Gryck-Pfeiler-Wagner

Aufgaben zur Matrizenrechnung und Linearen Optimierung

mit ausführlichen Lösungswegen.

1968. 309 Seiten. DM 16,— ISBN 3 7908 0024 4

Bloech, Jürgen, und Gösta-Bernd Ihde

Betriebliche Distributionsplanung

Zur Optimierung der logistischen Prozesse.

1972. ca. 120 S. ca. DM 20,— ISBN 3 7908 0109 7

Brauer, Karl M., und Mitarbeiter

Allgemeine Betriebswirtschaftslehre

Anleitungen zum Grundstudium mit Aufgaben, Übungsfällen und Lösungshinweisen.

2. Aufl. 1972. 404 S. DM 24,— ISBN 3 7908 0106 2

Hax, Herbert

Investitionstheorie

2. Aufl. 1972. 161 Seiten. DM 22,—

ISBN 3 7908 0108 9

Huch, Burkhard

Einführung in die Kostenrechnung

1971. 172 Seiten. DM 17,— ISBN 3 7908 0100 3

Lücke, Wolfgang

Produktions- und Kostentheorie

2. Aufl. 1970. 368 Seiten. DM 28,—

ISBN 3 7908 0005 8

Sasieni-Yaspan-Friedman

Methoden und Probleme der Unternehmensforschung

3. Nachdruck. 1971. 322 Seiten. DM 24,—

ISBN 3 7908 0025 2

Schneeweiss, Hans

Ökonometrie

1971. 384 Seiten. DM 36,— ISBN 3 7908 0008 2

Stenger, Horst

Stichprobentheorie

1971. 228 Seiten. DM 25,— ISBN 3 7908 0011 2

Swoboda, Peter

Finanzierungstheorie

in Vorb. ISBN 3 7908 0115 1

Vogt, Herbert

Einführung in die Wirtschaftsmathematik

1972. 240 Seiten. DM 20,—

ISBN 3 7908 0105 4

Vorobjoff, Nikolaj N.

Grundlagen der Spieltheorie und ihre praktische Bedeutung

2. Aufl. 1972. 84 Seiten. DM 9,—

ISBN 3 7908 0114 3

Preise: Stand Oktober 1972



physica-verlag · würzburg - wien